frontiers | Research Topics

Genomics of pathogens and vectors

Edited by
Tulio de Lima Campos, Gabriel Luz Wallau and
Augusto Simoes-Barbosa

Published in
Frontiers in Microbiology
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-5485-2
DOI 10.3389/978-2-8325-5485-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Genomics of pathogens and vectors

Topic editors

Tulio de Lima Campos — Aggeu Magalhães Institute (IAM), Brazil
Gabriel Luz Wallau — Aggeu Magalhães Institute (IAM), Brazil
Augusto Simoes-Barbosa — The University of Auckland, New Zealand

Citation

Campos, T. d. L., Wallau, G. L., Simoes-Barbosa, A., eds. (2024). *Genomics of pathogens and vectors*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5485-2

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 12 September 2024
DOI 10.3389/fgene.2024.1483676

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Ben Busby,
DNAnexus, Inc., United States

*CORRESPONDENCE
Túlio de Lima Campos,
✉ tulio.campos@fiocruz.br

†These authors have contributed equally to this work and share first authorship

RECEIVED 20 August 2024
ACCEPTED 05 September 2024
PUBLISHED 12 September 2024

CITATION
Wallau GL, Simoes-Barbosa A and Campos TdL (2024) Editorial: Genomics of pathogens and vectors.
Front. Genet. 15:1483676.
doi: 10.3389/fgene.2024.1483676

Editorial: Genomics of pathogens and vectors

Gabriel Luz Wallau[1,2,3†], Augusto Simoes-Barbosa[4†] and Túlio de Lima Campos[1*]

[1]Núcleo de Bioinformática, Instituto Aggeu Magalhães (IAM), Fundação Oswaldo Cruz (Fiocruz/PE), Recife, Brazil, [2]Departamento de Entomologia, Instituto Aggeu Magalhães (IAM), Fundação Oswaldo Cruz (Fiocruz/PE), Recife, Brazil, [3]Department of Arbovirology, Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Center for Arbovirus and Hemorrhagic Fever Reference and Research, National Reference Center for Tropical Infectious Diseases, Hamburg, Germany, [4]School of Biological Sciences, Faculty of Science, University of Auckland, Auckland, New Zealand

KEYWORDS

genomic surveillance, genomes, pathogens, antimicrobial resistance, virulence factors, diagnostics

Editorial on the Research Topic
Genomics of pathogens and vectors

Introduction

Advances in high throughput genomic sequencing and bioinformatics have revolutionized our understanding of the biology of pathogenic organisms, including those transmitted by vectors. The use of these technologies has enabled precise identification and tracking of pathogens based on mutational profiles. Elucidating transmission dynamics and evolutionary patterns is crucial for informing public health decision making, and this has become even more critical in our modern times as the accelerated climate change is rapidly altering pathogen transmission dynamics. Genomics has also been crucial to monitor antimicrobial resistance genes and other molecular aspects that can impair treatment efficacy. Furthermore, these methods have uncovered virulence factors, providing insights into mechanisms that enable pathogens to disperse in the environment, invade their hosts and cause disease. The information revealed by pathogen genomics is also pivotal for developing accurate diagnostics, effective vaccines, and targeted therapeutics thus becoming indispensable for advancing public health strategies aimed at tackling infectious diseases. With this in mind, the thematic issue "*Genomics of Pathogens and Vectors*" highlighted the use of genomics and bioinformatics to investigate a diversity of established and emerging pathogens. Together, these articles showcase how the application of these technologies contributes to advancing our knowledge on pathogen dispersion and transmission, resistance to treatment, virulence and pathogenicity, as well as diagnostics.

Enhancing molecular epidemiology

Genomics enables and enhances molecular epidemiology by providing detailed insights into the genetic factors that influence disease patterns and distribution within populations. For example, Chem et al. analysed genetic data from dengue viruses collected in Malaysia from 2015 to 2021. They found a high diversity of dengue viruses and frequent turnover of different strains, suggesting extensive movement of the virus between countries and regions. Wang et al. identified mutations in key genes associated with increased transmission of *Mycobacterium tuberculosis*, the causative agent of tuberculosis and one of the deadliest infectious disease worldwide. Focusing on population genomics of *Klebsiella pneumoniae* from China, Feng et al. showed a high prevalence of strains resistant to multiple antibiotics that may undergo rapid spread. This is a concerning bacterial pathogen in hospital settings, presenting a mortality rate of up to 40% in bloodstream infections. By sequencing the genomes of *Burkholderia pseudomallei* from environmental samples in Ghana, Schully et al. provide a better understanding on the distribution of this emerging pathogenic bacterium that causes meliodosis, a serious but neglected illness.

Monitoring antimicrobial resistance genes and virulence factors

Revealing genes that trigger antimicrobial resistance and virulence factors is crucial for understanding and combating infectious diseases. To understand the genetic relatedness between growth rate and disease severity, Zhu et al. compared the genomes of different *Mycobacterium* species and found that loss of virulence factors is central to this relationship. Following a similar comparative genomics approach, Flores-Oropeza et al. found a variety of genes associated with antibiotic resistance in *Escherichia coli* isolated from women with recurrent urinary tract infections. The researchers also identified a variety of genetic factors and phenotypic variations implicated on this disease. Another study by Cai et al. compared the genomes of the emerging human pathogen *Paraclostridium bifermentans* to a related bacterium species *Paeniclostridium sordellii*, revealing that the former has a larger and more plastic genome than expected and may encode more virulence factors than the reference species.

Studies in commensal and environmental species were also featured. In one example, Chen et al. analysed the genomes of *Cutibacterium granulosum*, a bacterium found on human skin. They found a high diversity of strains and identified genes potentially linked to antibiotic resistance and molecules linked to virulence. Yuan et al. analysed the genomes of *Ralstonia pickett*i, a bacterium found in environmental soil and water which can opportunistically cause human infections. These authors described a flexible genome that allows adaptation to different environments as well as antimicrobial resistance genes. Xiao et al. sequenced the genome of *Ralstonia solanacearum*, a bacterium causing a wilt disease in tobacco. They found high similarity across strains but also unique features on the phytopathogenic strain which may help explain how the latter infects tobacco and other plants. Lastly, Mpeyako et al. analysed genomes of *Trichomonas tenax* and compared specific gene sets between this species and *Trichomonas vaginalis,* two mucosal-dwelling protozoans of humans. While the latter is recognized as a true urogenital pathogen, the former has only been linked to gum disease. The authors identified genes linked to virulence across the two species that are involved on interactions with host cells and mucosal microbiota, supporting the notion that *T. tenax* may be a direct causative agent of gum disease.

Advancing molecular diagnostics

The integration of cutting-edge genomic and bioinformatic technologies enable timely detection of pathogens with improved precision and sensitivity. In this context, Yao et al. assessed the value of shotgun metagenomics for diagnosing respiratory tract infections. The authors showed that the metagenomic approach was more sensitive and accurate than traditional methods that rely on hypothesis-driven detection or pathogen isolation, especially for diagnosing acute infections. To overcome issues with growth and isolation of the fastidious pathogen *Francisella tularensis*, Isidro et al. developed a new method to capture and sequence its genome directly from animal samples. They successfully sequenced many full genomes, including mixed genotypes from single samples. This method will be useful for studying the spread of this zoonotic bacterium in wildlife and may be applicable to other pathogens and settings where microbial isolation is technically challenging.

Conclusion

In summary, the Research Topic of articles selected for this thematic issue showcases the transformative potential of genomics in advancing our battle against infectious diseases. Evolution, emergence and re-emergence of pathogens are reported with the increasing threats of antimicrobial resistance and hypervirulence. Meanwhile, recent and ongoing anthropogenic impact is changing the dynamics of emerging and reemerging pathogen transmission across the globe. Together, this underscores the urgency for expanding research, development and application of genomics and related technologies to safeguard public health. Genomics serves as a powerful tool for pathogen surveillance, diagnosis, and development of effective countermeasures. Equitable access to these technologies will strengthen global health security while fostering scientific growth, collaboration and innovation.

Author contributions

GW: Writing–review and editing. AS-B: Writing–review and editing. TC: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. GW is supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through their productivity research fellowships (310184/2023-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 09 November 2023
DOI 10.3389/fmicb.2023.1293206

Populational genomic insights of *Paraclostridium bifermentans* as an emerging human pathogen

Xunchao Cai[1,2†], Yao Peng[1†], Gongli Yang[1], Lijuan Feng[1], Xiaojuan Tian[1], Ping Huang[1], Yanping Mao[3]* and Long Xu[1,2]*

[1]Department of Gastroenterology and Hepatology, Shenzhen University General Hospital, Shenzhen University, Shenzhen, Guangdong, China, [2]Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, Guangdong, China, [3]College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen, Guangdong, China

Paraclostridium bifermentans (P.b) is an emerging human pathogen that is phylogenomically close to *Paeniclostridium sordellii* (P.s), while their populational genomic features and virulence capacity remain understudied. Here, we performed comparative genomic analyses of P.b and compared their pan-genomic features and virulence coding profiles to those of P.s. Our results revealed that P.b has a more plastic pangenome, a larger genome size, and a higher GC content than P.s. Interestingly, the P.b and P.s share similar core-genomic functions, but P.b encodes more functions in nutrient metabolism and energy conversion and fewer functions in host defense in their accessory-genomes. The P.b may initiate extracellular infection processes similar to those of P.s and *Clostridium perfringens* by encoding three toxin homologs (i.e., microbial collagenase, thiol-activated cytolysin, phospholipase C, which are involved in extracellular matrices degradation and membrane damaging) in their core-genomes. However, P.b is less toxic than the P.s by encoding fewer secretion toxins in the core-genome and fewer lethal toxins in the accessory-genome. Notably, P.b carries more toxins genes in their accessory-genomes, particularly those of plasmid origin. Moreover, three within-species and highly conserved plasmid groups, encoding virulence, gene acquisition, and adaptation, were carried by 25–33% of P.b strains and clustered by isolation source rather than geography. This study characterized the pan-genomic virulence features of P.b for the first time, and revealed that *P. bifermentans* is an emerging pathogen that can threaten human health in many aspects, emphasizing the importance of phenotypic and genomic characterizations of *in situ* clinical isolates.

KEYWORDS

emerging pathogen, clostridial toxins, comparative genomics, *Paraclostridium bifermentans*, virulence factors, conserved plasmids

Highlights

- P.b and P.s are hypothesized have similar mechanisms to initiate infections.
- P.b is probably less toxic than P.s by encoding fewer lethal toxins in the pangenome.
- Within-species-conserved and environmental-coevolved plasmids were identified in P.b.

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Leonardo Gabriel Panunzi,
CEA Saclay, France
Lara Ferrero Gómez,
Universidade Jean Piaget de Cabo, Cabo Verde

*CORRESPONDENCE
Yanping Mao
✉ maoy@szu.edu.cn
Long Xu
✉ longxu1012@szu.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 12 September 2023
ACCEPTED 26 October 2023
PUBLISHED 09 November 2023

CITATION
Cai X, Peng Y, Yang G, Feng L, Tian X, Huang P, Mao Y and Xu L (2023) Populational genomic insights of *Paraclostridium bifermentans* as an emerging human pathogen.
Front. Microbiol. 14:1293206.
doi: 10.3389/fmicb.2023.1293206

Introduction

Clostridia are bacteria belonging to the phylum *Firmicutes*, which are composed of a broad spectrum of Gram-positive (mostly), low GC content, spore-forming, and anaerobic bacilli (Orrell and Melnyk, 2021). Several genera of clostridia can cause mild to life-threatening diseases in both humans and animals, including tetanus and botulism, uterine infections, histotoxic infections and enteric diseases, by producing an array of protein toxins (Revitt-Mills et al., 2019; Zerrouki et al., 2021). Among the pathogenic clostridia, those grouped within clostridial cluster XI, such as *Clostridioides difficile* and *Paeniclostridium sordellii*, are particularly prevalent (Ezaki, 2009; Moore and Lacey, 2019). *C. difficile*, in particular, has attracted broad research efforts due to its clinical importance, and extensive studies have been conducted to clarify its virulence, strain diversity, transmission, and evolution at both the genomic and phenotypic levels (Knight et al., 2017). However, research on other clostridial cluster XI species, such as *Paraclostridium bifermentans*, whose clinical cases are uncommon, and has few genomic sequences publicly available or in most cases does not cause lethal diseases, is limited.

P. bifermentans is one of the two validly published species under the genus *Paraclostridium*, ubiquitously residing in various mesophilic conditions including soil, marine environments, polluted waters, and human bodies (Jyothsna et al., 2016). Our recent study has revealed that the well-characterized pathogen *P. sordellii* is phylogenomically closer to *P. bifermentans* than other pathogenic members within cluster XI (Zhao et al., 2022). *P. bifermentans* has traditionally been recognized as a human commensal, since it is usually nonpathogenic unless coexisting with *C. perfringens* (Weinberg and Séguin, 1918). However, a growing number of cases have been reported in clinical settings, demonstrating its ability to cause various human infections, such as brain abscess, lymphadenitis, necrotizing endometritis, joint infection, empyema, and endocarditis, particularly with the development of advanced diagnostic methods (Edagiz et al., 2015; Hale et al., 2016; Biswas et al., 2018; Barrett et al., 2020). Moreover, *P. bifermentans* PAGU1678^T was reported to exacerbate the pathological conditions of a dextran sulfate sodium-induced (DSS-induced) colitis mouse model in a recent study (Kutsuna et al., 2019). These findings have led us to appreciate that *P. bifermentans* has emerged as a pathogen in humans under specific conditions. While current observations suggest that infections caused by *P. bifermentans* are non-lethal, the possibility of high toxic strains cannot be excluded with such a limited number of reported cases.

As a member of clostridial cluster XI that is phylogenomically close to *P. bifermentans*, *P. sordellii* is commonly found in the rectal or vaginal tract of 3–4% of women, with the majority of carriers remaining asymptomatic (Aldape et al., 2016; Chong et al., 2016). Nevertheless, when pathogenic *P. sordellii* infections occur, they can rapidly progress and are associated with high mortality rates (~70%) due to the production of the lethal toxin protein TcsL, a member of the large clostridial toxin (LCT) family (Lee et al., 2020). It is worth noting that *C. difficile* also produces a toxin of the LCT family, namely TcdB, sharing 90% sequence identity with TcsL, which a major virulence factor responsible for *C. difficile* infections (Chen et al., 2018). In addition, clostridial species may also encode other potent virulence factors such as pore-forming cytotoxins, phospholipases, and metalloproteases that mediate their infections, many of which are encoded on extrachromosomal virulence plasmids (Revitt-Mills et al., 2019). Up to date, the virulence factor coding capability of *P. bifermentans* still remains largely unexplored, and as a result, our understanding of its pathogenesis and potential to cause lethal infections is limited.

In this study, we obtained the whole genome sequences of *P. bifermentans* and *P. sordellii* from the NCBI genome database. We analyzed the phylogenetic relationship between the two species by constructing a phylogenomic tree and performing whole-genome-sequence-based average nucleotide identity (ANI) analysis. Furthermore, we annotated and compared the pangenome and coregenome features of the two species, including general genomic characteristics such as genomic size, GC content, and coding density, as well as global gene functions and virulence factor coding profiles. Additionally, we explored the extrachromosomal toxin coding capacity of *P. bifermentans* for the first time by predicting, annotating, and grouping the virulence plasmid sequences based on sequence identity. The toxicity potential and pathogenesis of *P. bifermentans* were then proposed and discussed in detail.

Materials and methods

Genome sequence collection and preprocessing

The whole genome sequence sets from *Paraclostridium* spp. (i.e., *P. bifermentans* and *P. benzoelyticum*) and *P. sordellii* were downloaded from the NCBI genome Refseq database. The isolation sources of each strain were collected from the metadata table downloaded from the Refseq database or by searching the literature. The genomic sequences were evaluated for completeness and contamination using CheckM v1.0.12 (Parks et al., 2015). The standard lineage workflow "lineage_wf" from CheckM was performed to summarize general sequence features such as completeness, contaminations, genomic size, GC contents, coding density, and predicted rRNAs. Genomic sequences with completeness less than 70% or contamination greater than 10% were eliminated from the sequence sets. A total of 27 *Paraclostridium* spp. genomic sequences comprising one *P. benzoelyticum* strain JC272 and 26 *P. bifermentans* strains, and 60 *P. sordellii* genomic sequences were finally collected for further analysis (Supplementary Table S1).

Whole-genome-sequence-based phylogeny

Each genome sequence in the genome sets was annotated using Prokka v1.14.5 with default parameters. Aligned core and accessory genome sequences of the genome sets were extracted separately from the resultant gff files using the Roary v3.11.2 pangenome pipeline (Page et al., 2015). The phylogenomic tree of each alignment (i.e., core and accessory) was constructed using FastTree v2.1.11 with the generalized time-reversible model. The average nucleotide identity (ANI) between all the collected sequences was computed using fastANI v1.33 (Jain et al., 2018), and the genome distance (GD) was calculated using the phylonium v1.2 (Klötzl and Haubold, 2020). The tanglegram between the core-genome tree and accessory-genome tree was visualized using the dendextend v1.15.2 R package. The ANI matrix and GD matrix were visualized using the pheatmap v1.0.12 R package. Finally, the core-genome tree with general features was visualized using iTol v6 (Letunic and Bork, 2021).

Pan-genomic features and functional genome annotation

Plasmid sequences in the genome sets were predicted using (1) PlasForest v1.2,[1] a machine learning-based tool that uses sequence homology, and (2) BLAST-based method with an e-value lower than $1e^{-5}$ and a sequence similarity higher than 90% and query coverage higher than 50% by aligning to the documented plasmids from *Paraclostridium* spp., *P. sordellii* and *C. perfringens*, and (3) circular contigs with no *dnaA* gene and no rRNA coding sequences although not predicted as plasmids based on the sequence homology methods. Pangenome features of either *P. bifermentans* or *P. sordellii* were extracted using the Roary v3.11.2 pangenome pipeline from the gff files produced by Prokka v1.14.5. Subsequently, the extracted pangenome features were analyzed and visualized using the R package Pagoo v0.3.12 (Ferrés and Iraola, 2021). The core genome was defined as the set of CDSs that are present in over 95% of genomes. All other CDSs were classified as the accessory-genome, which was further subdivided into the cloud genome (i.e., those present in less than 5% of genomes) and the shell genome (i.e., those present in 5–95% of genomes). The functional genome annotation was conducted using eggNOG-mapper v2.1.5, which comprehensively annotated the Carbohydrate-Active enZYmes (CAZy), Cluster of Orthologous Groups (COG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Cantalapiedra et al., 2021). Virulence factor coding genes, including toxin genes and antimicrobial resistant genes, in the pangenome were predicted using PathoFact v1.0 with the "complete" parameter (de Nies et al., 2021).

Results

General genomic features and phylogeny characterization of *Paraclostridium bifermentans*

Prior to characterizing the genomic features, we carefully checked the isolation source of the strains. Interestingly, despite the ubiquitous presence of *P. sordellii* strains in the environment, none of the strains that had genomes sequenced are isolated from the environment, while 15 originated from animal sources and 45 from human sources. In contrast, strain *P. benzoelyticum* JC272 was isolated from the environment, and 6, 6, and 13 out of the 26 *P. bifermentans* strains were isolated from environmental, animal, and human sources, respectively (Figure 1A). Our analysis of the whole genome tree for both *P. bifermentans* (P.b) and *P. sordellii* (P.s) did not reveal any correlation between isolation source and within-species phylogeny, indicating that isolation source did not impact the phylogenetic relationships within these species (Figure 1A). Then, the collected genome assemblies were compared on assembly level, genome size, GC content, completeness, contamination, and predicted genes (Figure 1A). All of the genomes showed high completeness (> 99%), 4 out of the 87 genomes had medium contamination (5% < contamination <10%), and only 1 was assembled into more than 500 scaffolds. These features suggest that the genome set is of high quality. However, only 4 genomes were assembled at the completed level (2 from P.b and 2 from P.s), and 3 at the chromosomal level, and the others at the scaffold or contigs levels (Figure 1A). Of note, the rRNA sequence numbers showed high variability between strains within and between species. This is likely due to the limitations of short-read sequencing and assembly methods in accurately assembling the multicopy rRNA genes in bacteria. Analysis of completed genomes revealed that P.b and P.s possess a high number of rRNA genes, ranging from 48 to 51, which suggests a high protein synthesis capacity for these two species (Figure 1A). Additionally, these two species showed similar coding density, ranging from 0.84 ~ 0.87 (Figure 1B). The genome size and GC content of P.b are significantly higher than those of P.s (Figure 1B). For the following reasons, five genome sequences were excluded from further functional analyses: (1) they were identified as outliers in genome size and were assembled from metagenomes (i.e., MAG072, L3_131_244G1_dasL3_131_244G1_concoct_40_sub); (2) they were identified as belonging to a different species (i.e., strain JC272 of *P. benzoelyticum*); (3) they were identified as outliers in both genome size and GC content (i.e., BIOML-A1, and BIOML-A2) (Figure 1B).

As shown by the core-genome tree constructed from P.b and P.s, *P. bifermentans* is significantly divergent from *P. sordellii* (Supplementary Figure S1). The tangled dendrogram of the trees is displayed in Figure 2A. In detail, P.b and P.s are clearly separated by the topological structure of both the core-genomic and accessory-genomic tree, from which the tanglegram displayed highly consistence (entanglement = 0.11), underscoring a distinct divergence in phylogeny between P.b and P.s across both the core-genome and accessory-genome (Figure 2A). Furthermore, the heatmap and hierarchical clustering from ANI and GD further reveal the phylogenomic divergence of P.b and P.s at the species level (Figures 2B,C), highlighting a robust separation between the two species and substantiating the evolutionary divergence at both the core and accessory genomic levels.

Pangenome characterization of *Paraclostridium bifermentans*

To compare the coding capacity of P.b and P.s, we performed pangenome analyses using Roary. The pangenome of 24 *P. bifermentans* genomes and 58 *P. sordellii* genomes consisted of 10,030 and 11,076 orthologous genes, respectively. The pangenome curves for both species formed a "C" shape, and the alpha index of the heaps law models being lower than 1 (Figure 3A) (Tettelin et al., 2008), which revealed that both P.b and P.s carried open pangenomes (Figure 3A). The pangenome curves of P.b and P.s were best fitted for the positive-power model, with the formulas $yP.b = 3284.7\chi 0.3451\left(R^2 = 0.9971\right)$ and $yP.s = 2948.1\chi 0.3153\left(\mathrm{R}^2 = 0.9915\right)$, respectively, indicating that the pangenome of P.b is more plastic than that of P.s ($yP.b > yP.s$) (Figure 3A). Moreover, the core-genome size of P.b (2384) is comparatively smaller than that of the P.s (2526) (Figures 3B,C),

[1] https://github.com/leaemiliepradier/PlasForest

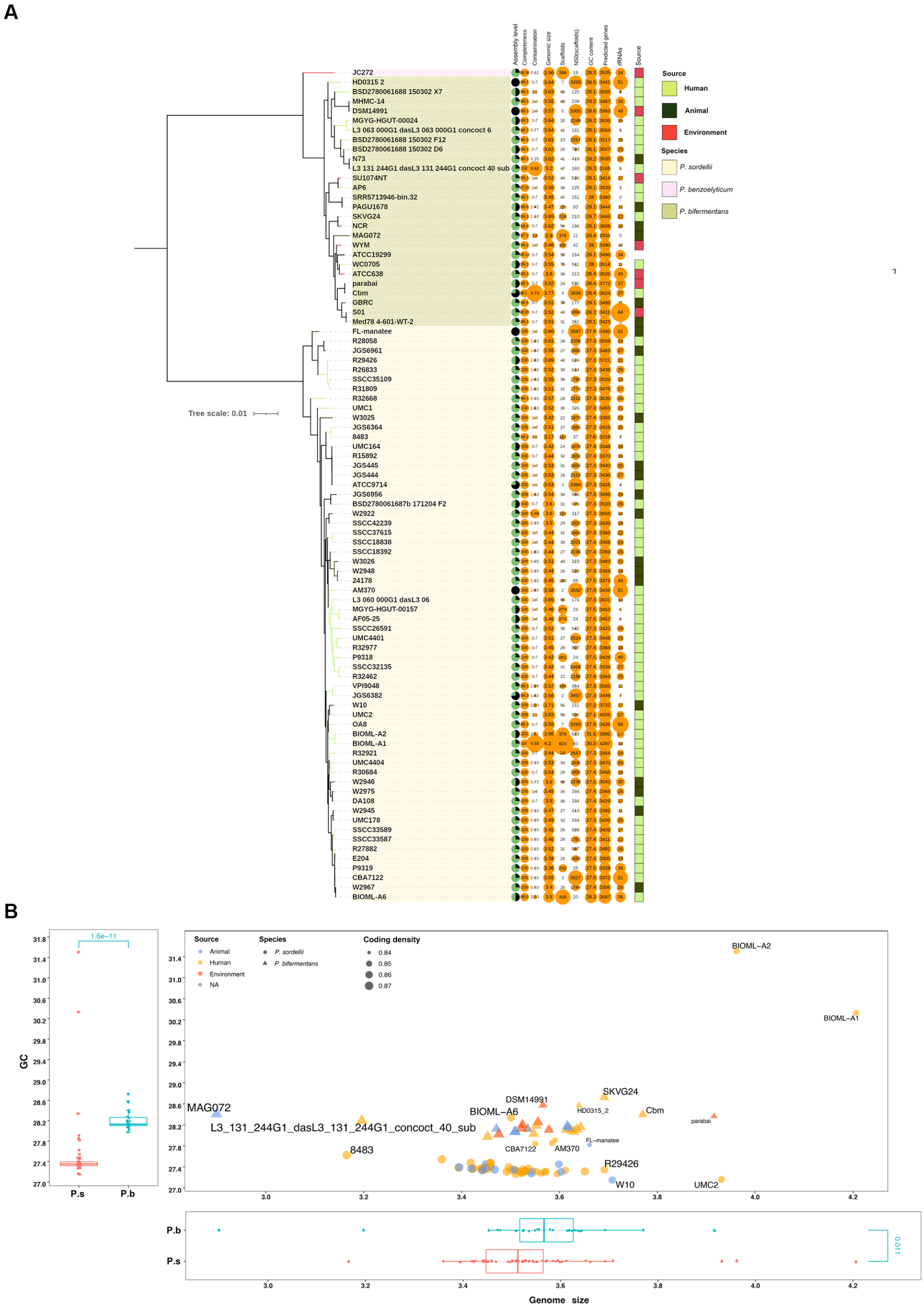


FIGURE 1
General genomic features of *P. bifermentans* and *P. sordellii*. **(A)** General genomic features and isolation sources of *P. bifermentans* and *P. sordellii* strains. The blank strip in the right means an undetermined isolation source. **(B)** Comparison of genomic size and GC content between *P. bifermentans* and *P. sordellii*. P.b and P. s represent *P. bifermentans* and *P. sordellii*, respectively. Differences in GC content and genomic size between P.b and P.s were analyzed using the Wilcox Test, with $p < 0.05$ considered statistically significant.

which is further corroborated by the core-genome curves (Figure 3A). As the average genomic size of P.b is larger than that of P.s (Figure 2), the comparison in core-genome size between P.b and P.s further supported that P.b had a more plastic pangenome than that of P.s by carrying a greater number of accessory CDSs. The core-genome curve showed that 1949 orthologous genes are shared by all P.b strains, and 1912 orthologous genes are shared by all P.s strains, with the strain-specific genes in P.b and P.s are 3,673 and 4,095, respectively (Figures 3A–D). The core-genomes of P.b and P.s accounted for only 23.77% (2,384/10030) and 22.81% (2,526/11076) of the respective pangenome (Figures 3B,C). The core-genome-coded COG functions between P.b and P.s were similar in COG

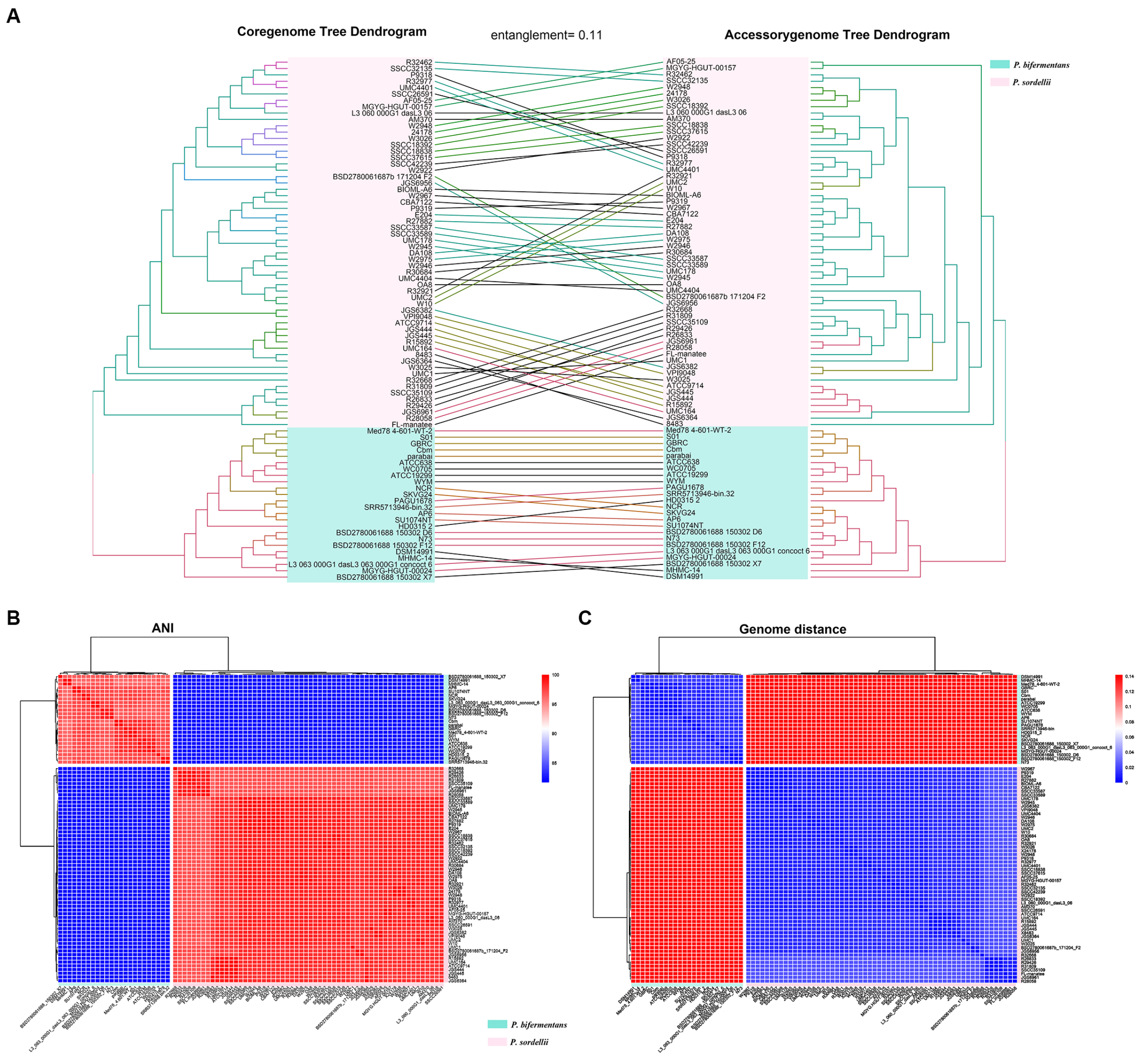


FIGURE 2
Phylogeny relationships between *P. bifermentans* and *P. sordellii* at the whole-genome sequence level. **(A)** Phylogenomic tree constructed using the core genome and accessory genome of the *P. bifermentans* spp. and *P. sordellii* genomes set. **(B)** ANI analysis between *P. bifermentans* and *P. sordellii* based on whole genome sequences. **(C)** Genome distances between *P. bifermentans* and *P. sordellii* based on whole-genome sequences.

categories and counts, while the accessory-genome-coded COG functions showed high differences by COG counts (Figure 3D). P.b's accessory-genome, for example, encoded more COG functions in nutrient metabolism, protein synthesis and energy conversion (e.g., "C Energy production and conversion," "G Carbohydrate transport and metabolism," "I Lipid transport and metabolism," "J Translation, ribosomal structure and biogenesis," "K Transcription," "O Posttranslational modification, protein turnover, chaperones," "P Inorganic ion transport and metabolism," "T Signal transduction mechanisms" and "U Intracellular trafficking, secretion, and vesicular transport"), while that of P.s encoded more COG functions in cell cycle and host adaptation (e.g., "D Cell cycle control, cell division, chromosome partitioning," "L Replication, recombination and repair," "M Cell wall/membrane/envelope biogenesis," "N Cell motility," "S Function unknown" and "V Defense mechanisms") (Figure 3D).

Virulence genes carried by *Paraclostridium bifermentans*

We further explored the virulence factor coding capacity of P.b and P.s to compare their pathogenic potentials. The virulence gene profiles between P.b and P.s were divergent, enabling effective discrimination between the two species (Supplementary Figure S2a). Specifically, each P.s strain carried an average of 71 ± 3 virulence genes, while each P.b strain carried an average of 70 ± 4 virulence genes. A total of 176 and 167 virulence genes were carried by the pangenomes of P.s and P.b, respectively. Among them, 29.34% (49/167) and 31.25% (55/176) belonged to the core-genomes of P.b and P.s, respectively, accounting for approximately 70% of the virulence genes in each strain (Supplementary Figure S2B, Figure 4A). Furthermore, P.s carried 39 virulence genes encoding secreted toxins in the pangenome, and 14 of them were in the core-genome, whereas that of P.b was 28,

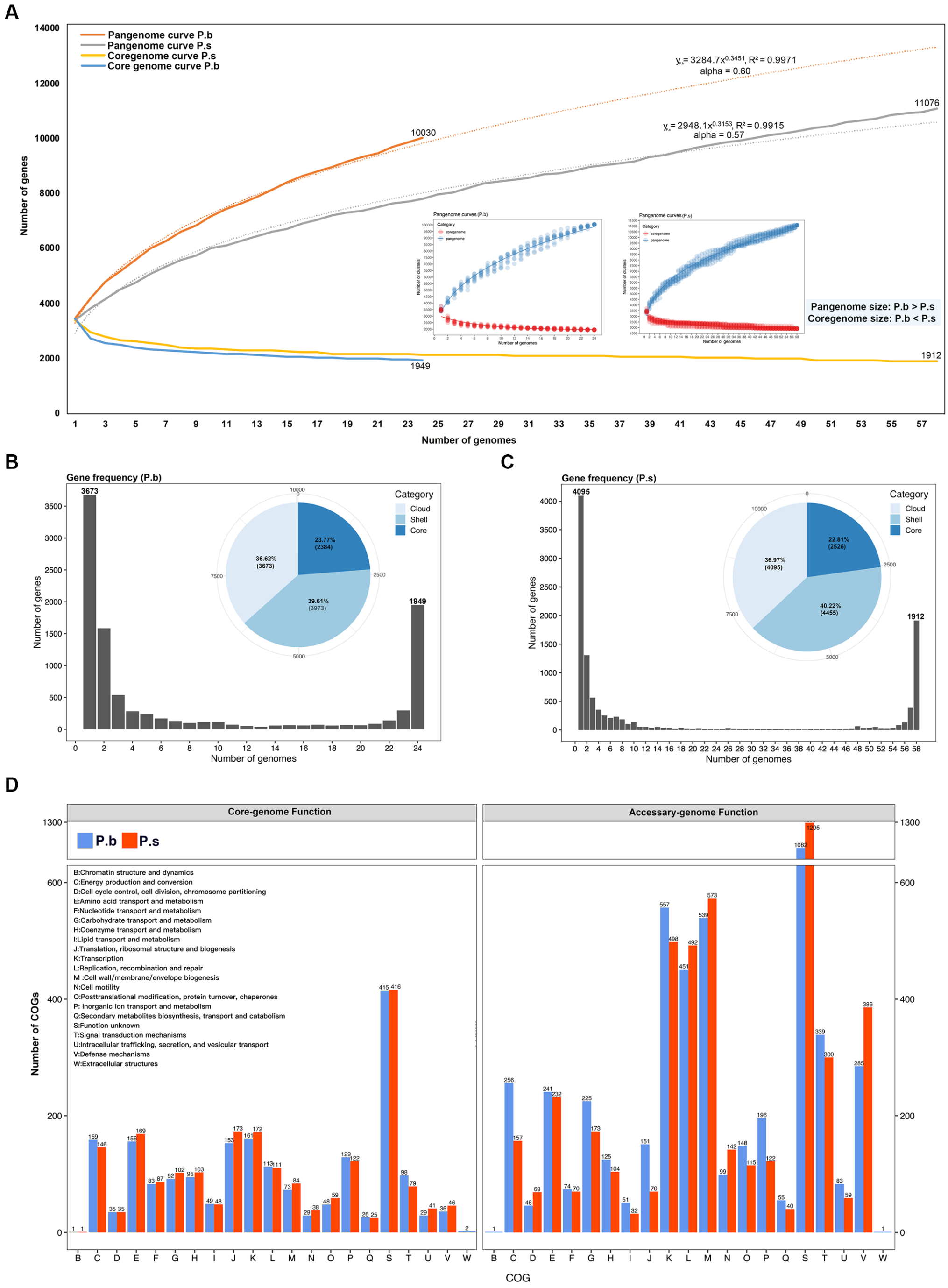


FIGURE 3
Comparison of pan-genomic features between *P. bifermentans* and *P. sordellii* strains. **(A)** Pan-genomic and core-genomic curves fitted using the positive-power model. **(B,C)** Gene frequency histogram and pangenome category pie-chart. **(D)** Pan-genomic functions displayed by clusters of orthologous groups (COGs).

and only 9 of them were in the core-genome. Since we had confirmed in this study that P.b carried a much opener pangenome than that of P.s (Figure 3A), we can deduct from the results here that the core-genome of P.b may encode less secreted-toxin genes than that of P.s, with the available P.b genomes increased. In addition, 44.91% (75/167) of the virulence genes in P.b were from MGEs, 19 of which are core genes, and over half of the MGE-related virulence genes (40/75) are phage-originated (Figure 4A). Similar findings were observed in P.s (Supplementary Figure S2B). The MGEs, especially the phage-originated MGEs, play vital roles in the expansion of the virulence gene pool in the pangenomes of P.b and P.s (Figure 4A, Supplementary Figure S2B). Regarding virulence gene families, most of them were shared by the pangenomes of P.b and P.s, but the gene counts varied. For example, the P.b pangenome coded more Zinc-dependent phospholipase C, UDP-glucose 4-epimerase C-term subunit, PLD-like domain, and Nitroreductase family virulence

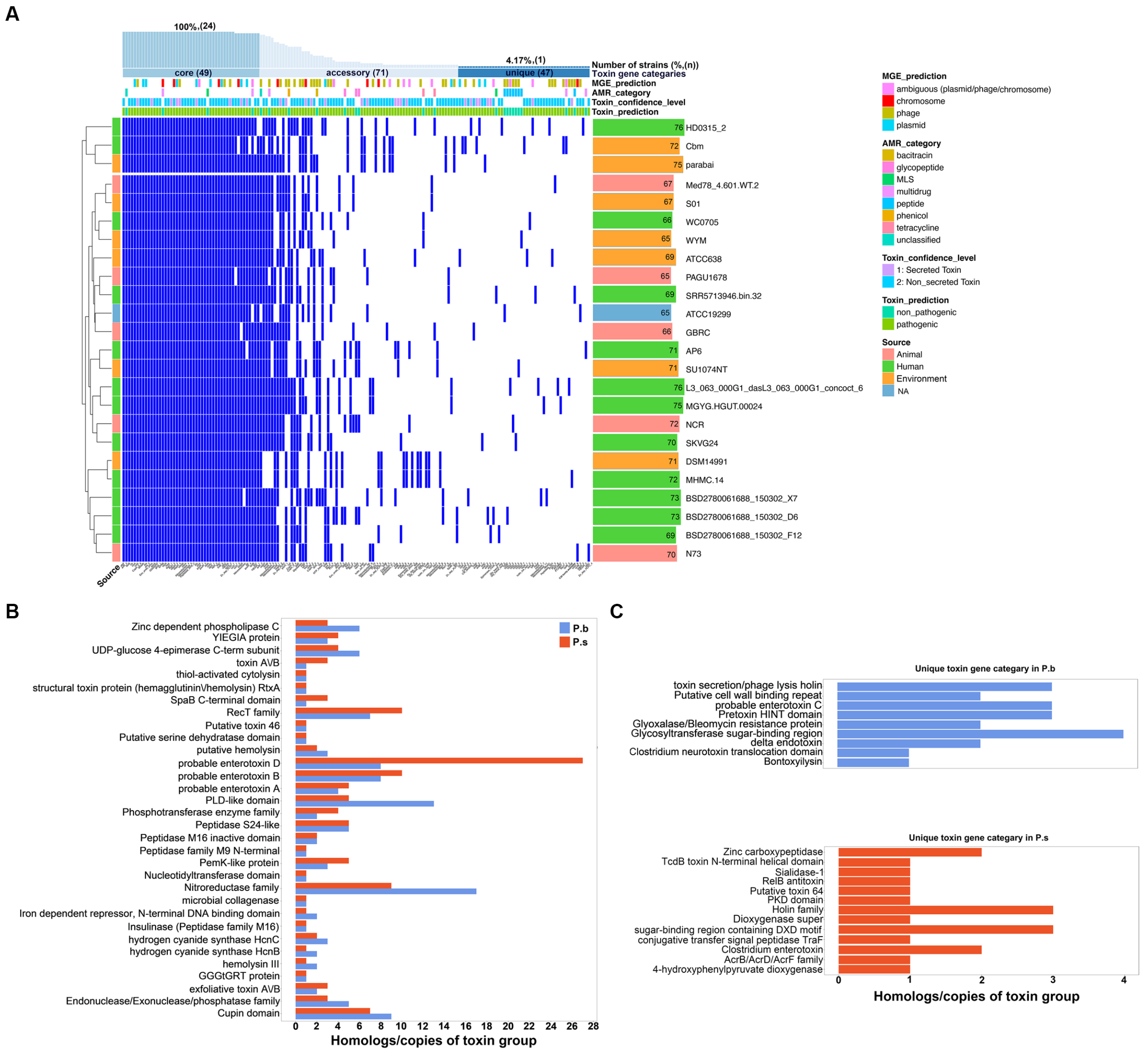


FIGURE 4
Characterization of the virulence factor coding capacity of *P. bifermentans*. **(A)** Heatmap showing the presence or absence of virulence factor coding genes in the pangenome of *P. bifermentans*. **(B,C)** The shared **(B)** or unique **(C)** virulence gene categories between *P. bifermentans* and *P. sordellii*.

factors, while that of P.s coded more toxin A/B, SpaB C-terminal domain, and enterotoxin D virulence factors (Figure 4B). Additionally, there are several species-specific toxin gene families, such as enterotoxin C, pretoxin HINT domain, and Clostridium neurotoxin translocation domain in P.b, and Zinc carboxypeptidase, and Clostridium enterotoxin in P.s (Figure 4C). A detailed virulence gene co-occurrence network between P.b and P.s is shown in Supplementary Figure S3.

We further compared the toxin gene counts derived from different sequence types to estimate the virulence coding capacity of P.b and P.s. The results showed no significant difference in the pangenome-coded toxin gene counts between P.b and P.s. However, the coregenome-coded toxin genes in P.s were higher than those in P.b (Figures 5A–C). Moreover, both chromosomal- and phage-originated toxin gene counts in the pangenome or coregenome of P.s were higher than that in P.b (Figures 5A,B), which displayed an opposite pattern in the accessory-genome (Figure 5C). It is noteworthy that the plasmid-originated toxin genes were higher in both the pangenome and coregenome of P.b than in P.s (Figures 5A,B). As a result, the plasmid-originated sequences played a significant role in shaping the toxin gene profile of P.b's coregenome, while the phage-originated sequences played a significant role in that of the accessory-genome (Figures 5B,C).

Furthermore, we compared the presence of the common reported clostridial toxin genes in P.b and P.s. The results showed that the pangenomes carried a total of eight toxin gene homologs, out of which five, namely *colA*, *pfo*, *plc*, *toxB/tcsL*, and *iap*, were shared by both species. The former three genes were core genes in both species that coded for proteins involved in host cell surface attachment and cell membrane damage, indicating that P.b and P.s may share similar mechanisms for initiating infections (Table 1). The *toxB/tcsL* and *iap* that encode host cell intracellular toxicity,

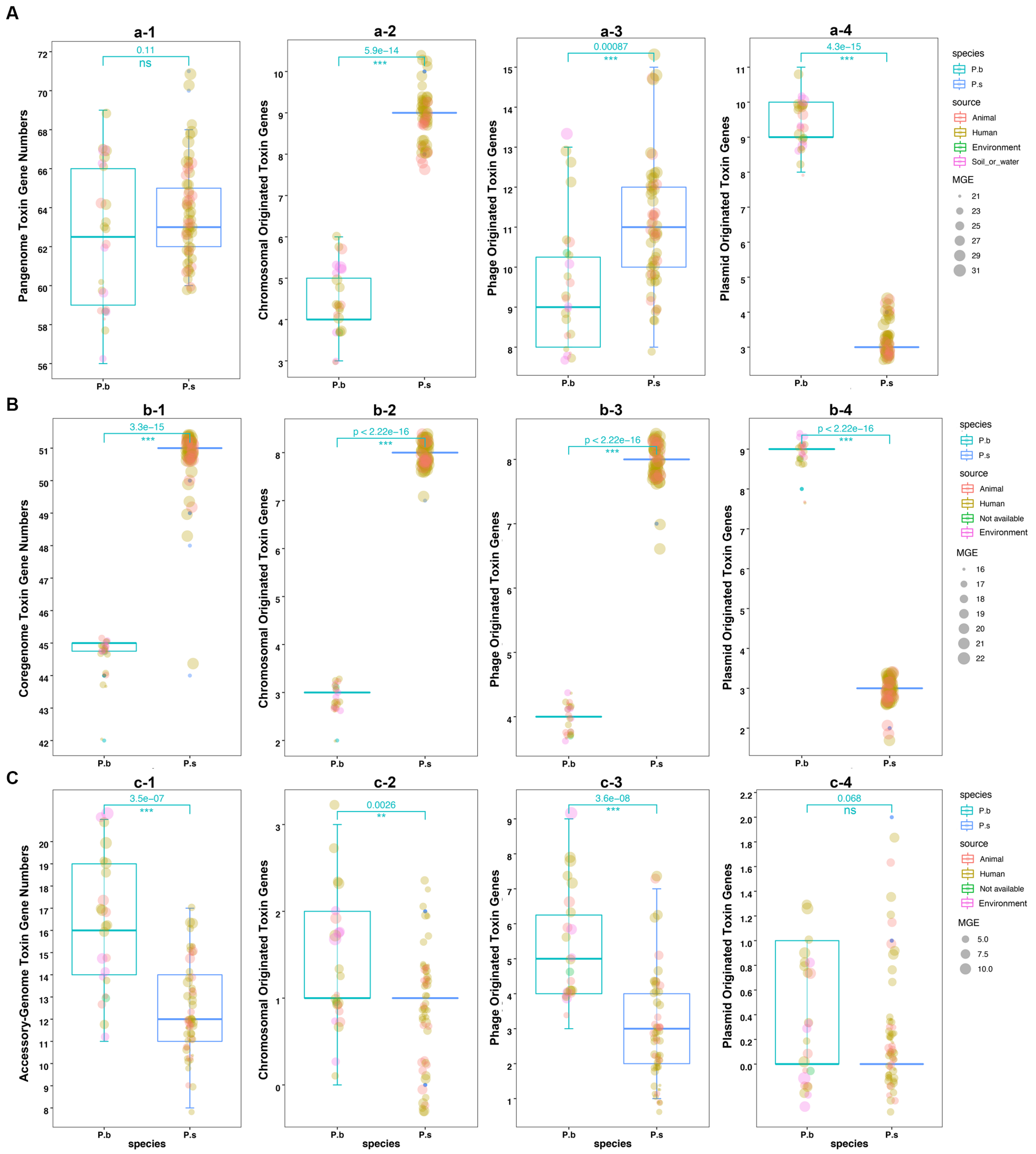


FIGURE 5
Comparison of the toxin genes in *P. bifermentans* and *P. sordellii* by sequence origination types. Toxin genes coded in pangenome **(A)**, coregenome **(B)**, and accessory-genome **(C)**. The subplots (e.g., a-2, a-3, a-4) indicate the toxin gene sequence type from the MGE prediction using PathoFact, and unclassified predictions were excluded from the plots. The Wilcox Test was used to compare the difference in toxin gene counts between the pangenome, coregenome, and accessory-genome, as well as different sequence-origination types, with a *p*-value <0.05 considered significant.

which may lead to lethal and dermonecrotic effects, are presented more frequently in P.s than in P.b, indicating that the P.s is more toxic than the P.b (Table 1). The remaining three genes, namely *toxA*/*tcsH*, *plcB* and *cpe*, were specifically carried by P.s, coding functions of hemorrhagic toxin, pathogen's vacuole escape, and cell pore-forming. Furthermore, the *plcB* gene was a core gene in P.s but absent in P.b (Table 1).

Virulence plasmid groups in *Paraclostridium bifermentans*

To elucidate the role of plasmids in the virulence of P. b, the plasmid sequences in the genomes of P.b were predicted and extracted using the methods described in the methods section. Plasmid pPbm14_8 from strain Cbm was depleted in further analyses because

TABLE 1 Presence of typical clostridial toxin gene homologs in *P. bifermentans* and *P. sordellii.*

Toxin gene homologs[a]	Presence in P.b (%, *n*/*n*)	Presence in P.s (%, *n*/*n*)	Gene products	Possible virulence mechanisms
colA	100%, 24/24	100%, 58/58	Microbial collagenase	Extracellular matrices degradation
pfo	100%, 24/24	98.28%, 57/58	PFO (thiol-activated cytolysin)	Pore-formation
plc	95.83%, 23/24	100%, 58/58	PLC (Zinc dependent phospholipase C)	Membrane damage
toxB/tcsL	4.17%, 1/24	12.07%, 7/58	toxin B	Infiltration and destroying
iap	12.5%, 3/24	100%, 58/58	Probable enterotoxin	Actin cytoskeleton destroying
toxA/tcsH	0	12.07%, 7/58	toxin A	Cytoskeletal structure disruption
plcB	0	100%, 58/58	Peptidase M16 inactive domain	Pathogen's vacuole escape mediation
cpe	0	3.45%, 2/58	Clostridium enterotoxin	Pore-formation

[a]Gene symbol was designated by using the Prokka and the gene function was predicted using the PathoFact.

of that the pPbm14_8 was identified as a chromosome sequence by aligning to the NCBI-nt database (Supplementary Table S2). By aligning the reported circular plasmid sequences in P.b and P.s strains, we then found that all plasmids between species are divergent, while some of those are highly conserved within-species (Supplementary Figure S4a). Further analyses on the plasmid sequence similarity between *C. perfringens*, P.b and P.s also revealed high within-species conservation (Supplementary Figure S4b). Finally, eight conserved plasmid groups in P.b were identified, ranging from 25 kb to198 kb in length (Figures 6A–F, Supplementary Figure S5). Among them, group_a, group_b and group_c are the most conserved groups with a presence in 25–33.33% P.b. strains, which were carried by 8, 8 and 6 P.b strains, respectively (Figures 6A–C). Plasmids belonging to group_a are the longest and two of them (i.e., pHD0315_2–1 and DSM14991 unnamed1) were assembled at the completed level (Figure 6A). The longest plasmid in group_a is from strain BSD_D6, the NODE_4 in the draft genome, which is 198,761 bp, 40 kb longer than other plasmids in this group (Figure 6A). This group of plasmids encode about 150–180 CDSs, of which one third can be annotated, and is the best annotated plasmid group in P.b (Figure 6A). The reference plasmid in this group carries eight potential virulence genes, of which five (i.e., *Cupin_2*, *nitroreductase family*, *puuR*, *corC_2*, and *tlyC*) are core genes in P.b (Figures 4A, 6A). It is noteworthy that the plasmids in this group are essential for regulating bacterial growth and adaptation through the encoding of various functions related to organic sulfate reduction (*asrABC*, *nirC*), 5,6-dimethylbenzimidazole synthesis (*nox_2*), styrene degradation (*gatA*), aminoacyl-tRNA biosynthesis (*aspC*, *gatA*), etc. (Figure 6A). Additionally, this group of plasmids encode five virulence genes, two of which, namely *cat* and *drrA*_2, confer resistance to chloramphenicol and doxorubicin. Group_f represents a rare but typical virulence plasmid group in P.b that is exclusively carried by two strains, namely parabai and Cbm, isolated from anopheline-endemic areas in Malaysia and Brazil, respectively. This plasmid group encodes two pesticidal toxins (Cry16Aa and Cry17Aa) and one novel neurotoxin, which are flanked by Tn3 and IS1182 family transposases, respectively (Figure 6F). Notably, the Tn3 transposase is also encoded by other plasmid groups (e.g., group_b and group_c) (Figures 6B,C). While group_b and group_e plasmids are less annotated, some of the genes they carry, such as *pqqD*, *srtB* and *dtpT*, are related to pyrroloquinoline quinone synthesis (Ludueña et al., 2017), surface anchoring (Weiss et al., 2004), and salt stress protection (Wouters et al., 2005). Conversely, group_d, group_g and group_h plasmids are poorly annotated (Figure 6D, Supplementary Figures S5g,h). Interestingly, strains carrying plasmids clustered in the same conservation group were isolated from distant geographic locations (Figure 6G), which may be clustered by the isolation-source and have co-evolved with their habitats, as we found that the group_a and group_b plasmids are predominantly carried by strains isolated from feces (Figure 6G). Furthermore, strains with completed genome sequences (i.e., HD0315_2 and DSM14991) carried more plasmids than those with draft genome sequences. Unfortunately, over half of the strains with genome sequences are not isolated from clinical settings (Figure 6G).

Discussion

As a species that is phylogenomically close to the pathogen P.s in clostridial XI, P.b is being recognized as an emerging human pathogen. Despite this, the strain-level diversity, toxicity, and genomic virulence capacity of P.b strains are yet understudied. This study reveals that P.b possesses a larger genome size with a higher GC content than P.s. Additionally, both P.b and P.s possess an open pangenome, with that of P.b being more plastic. While the gene function profiles of P.b and P.s are similar in the core-genomes, they differ significantly in the accessory genomes. P.s is more toxic than P.b, as it carries more toxin genes and encodes more secretion toxins in the core-genome and carries lethal toxin genes in the accessory genome. However, plasmid-originated toxin genes are more abundant in the pangenome or core-genome of P.b. Notably, plasmids were discovered to be species-conserved among P.b., P.s., and C.p. Furthermore, these plasmids were discovered to be clustered not by geographic location but by isolation source in P.b. This study also identified conserved Tn3 and IS1182 family transposase coding genes flanked by toxin coding genes in P.b, which may facilitate the acquisition and spread of virulence genes for P.b strains.

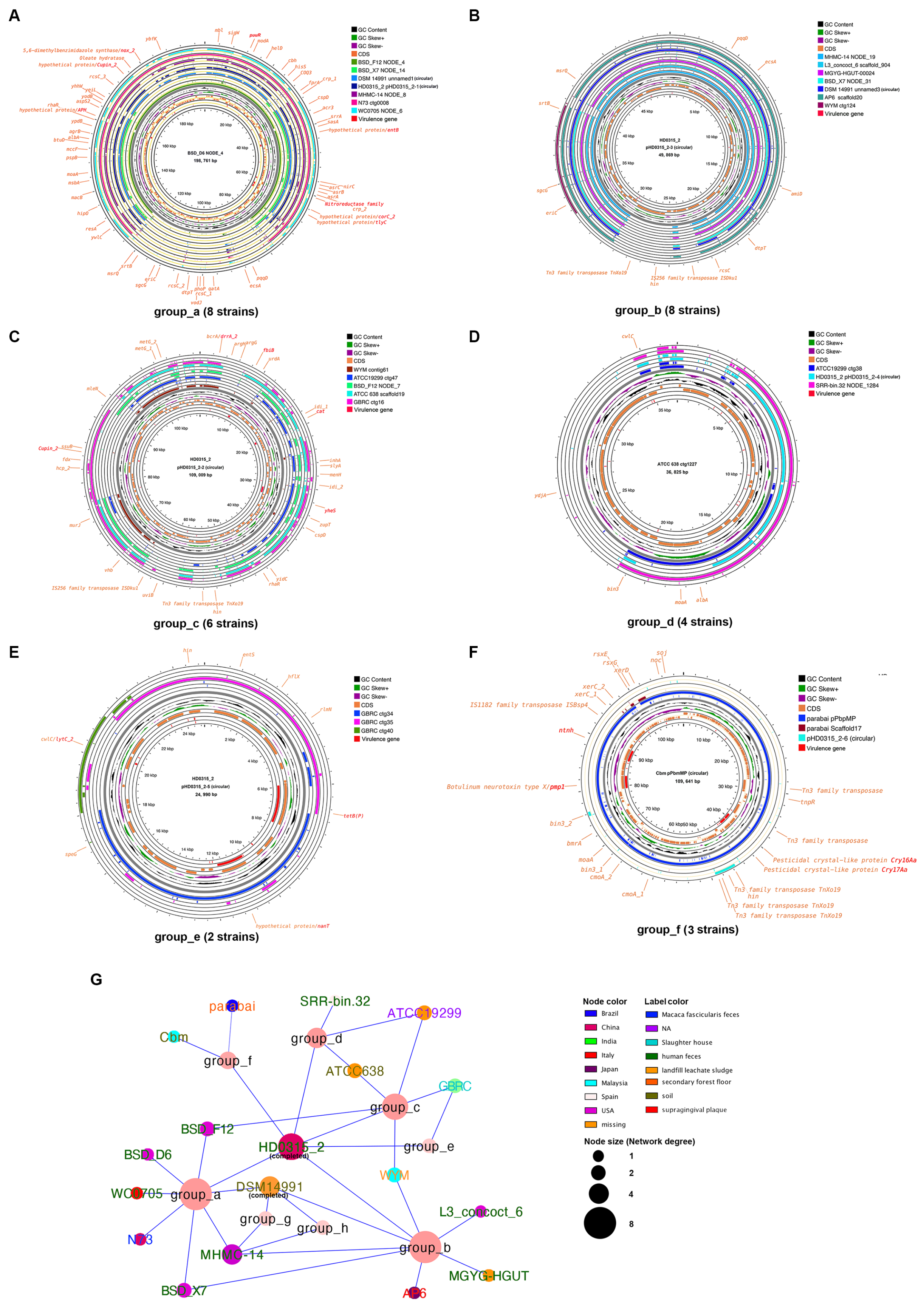


FIGURE 6 (Continued)

FIGURE 6 (Continued)
Representative plasmid groups in *P. bifermentans* and their coded functions. **(A–F)** Group_a to group_f plasmids in *P. bifermentans*. The longest plasmid sequence in each group was represented as the reference sequence. CDSs annotated as hypothetical proteins were not labeled in the plots. **(G)** Co-existence network of the conserved plasmid groups in P.b.

Previous studies have revealed that a larger genome and higher GC content elicit bacteria adaptation in more complex and varied environments (Lassalle et al., 2015). In the case of P.b, this species possesses a larger genomic size and higher GC content than P.s, which may confer an enhanced capacity to thrive across diverse environments (Figure 1). Our study revealed that P.b strains were isolated from a broader range of sources, despite the fact that the number of sequenced P.b genomes is currently fewer than those of P.s (Figure 1A). The versatility of P.b can be attributed to its adaptive capacity and genomic plasticity, which is primarily determined by the accessory genome. For instance, P.b strains have been utilized in various applications such as metal-contamination bioremediation (Neveling et al., 2022), antibiotic degradation (Fang et al., 2021) and organic compound conversion (Huang et al., 2022). In contrast, P.s strains, those that had genomes sequenced were all isolated from animal or human hosts (Figure 1A) and are commonly involved in infections (Zerrouki et al., 2021; Gonzalez-Astudillo et al., 2022), indicating that they may have adapted to these environments by encoding more host defense functions (Figure 3D). Nevertheless, it is important to note that the host adaptation capacity of P.b should not be underestimated, as the P.b carries a highly plastic pangenome, whereas the number of isolated P.b strains and sequenced genomes from clinical specimens is increasing.

P.b has been reported to be clustered into the clostridial cluster XI with *C. difficile* and P. s, and phylogenomically closer to the latter (Moore and Lacey, 2019; Zhao et al., 2022). The virulence of *C. difficile* and P. s is largely determined by their production of a wide array of toxins, although some non-toxin factors, such as degradative enzymes, surface-exposed proteins, and adhesion factors, also play a role (Revitt-Mills et al., 2019; Geier et al., 2021). We predicted for the first time the full-spectrum virulence gene profile of P.b, revealing a distinct virulence gene pool compared to that of P.s (Figure 4). P.s possesses a higher pathogenic potential than P.b, due to the encoding of more secreted toxins in the coregenome and more LCT toxins in the accessory genome (Figure 4). Although P.b infections are much rarer than P.s infections in clinical settings (Revitt-Mills et al., 2019), and no lethal infections have been reported, it is possible that lethal infections of P.b may occur in humans. This is supported by our identification of *toxB* and *iap* homologs in 4–12% of P.b genomes (Figure 4A).

It is worth noting that P. b and P. s may share similar extracellular infection mechanisms with C.p, as evidenced by the presence of three core toxin gene homologs (i.e., *colA*, *pfo* and *plc*) (Table 1), which have been reported to lead to extracellular matrix degradation, cell pore-formation and cell membrane damage in host cells (Popoff and Bouvet, 2009). Studies have demonstrated that these toxins can exhibit synergistic effects during the infections (Popoff and Bouvet, 2009; Popoff, 2016). Nevertheless, the intracellular cytotoxicity of P.b may be much lower than that of P.s, because only one toxin gene (i.e., *iap*), encoding actin cytoskeleton destroying functions, is present in 12.5% of the P.b strains compared to 100% present in P.s strains (Table 1). Additionally, P.s strains exclusively carried three toxin genes encoding cytoskeletal structure disruption, pathogen's vacuole escaping and pore-formation, making them more lethal than P.b strains (Table 1). Further sequence similarity analysis using NCBI blastp revealed that the three core toxin protein homologs clustered by a species-conserved pattern and were divergent from that of C.p (Supplementary Figure S6). Previous research has revealed that the P.b PLC is weakly hemolytic and nonlethal compared to the C.p PLC, and the C-terminal of P.b PLC lacks two residues, Tyr331 and Phe334, that are present in the C.p PLC (Popoff and Bouvet, 2009). Moreover, varying degrees of pore-formation cytotoxicity of the PFO homologs have also been observed in *Bacillus*, *Listeria*, and *Streptococcus* (Hotze et al., 2013), contributing to pathogenesis through processes such as intracellular protein translocation (*S. pyogenes*), phagosomal escaping (*L. monocytogenes*), and macrophage escaping (C.p) (Portnoy et al., 1988; Madden et al., 2001; O'Brien and Melville, 2004). However, the specific effects of PFO and PLC homologs in P.b and P.s on cytotoxicity are yet to be explored. As a matter of fact, the P.b and P.s strains rarely cause human infections, from which we can deduce that the PLC and PFO homologs in them are less toxic than those found in C.p. However, wet-bench experiments are required to clarify the exact cytotoxicity of these toxin homologs in the future.

Many clostridial virulence factors, particularly lethal toxins, are encoded on large plasmids or other MGEs (e.g., bacteriophage, insert elements, and transposons). The mobile nature of these virulence genes facilitates their wide dissemination within the clostridial genera (Revitt-Mills et al., 2019). The lethal LCTs and the associated regulators in P.s (i.e., TcsL and TcsH) have been found encoded by the pCS1 family plasmids, which are conserved in at least seven strains (Vidor et al., 2018). In addition, all members of the pCS1 family encode a number of surface-exposed proteins (e.g., sortase enzyme (*srtB*)). Other plasmid families, such as pCS2, have also been identified in P.s, but they are poorly characterized (Couchman et al., 2015). Our study demonstrated that the number of plasmid-originated toxin genes in P.b is greater than that in P.s in the coregenome (Figure 5), and eight conserved plasmid groups were identified in P.b (Figure 6). Interestingly, no inter-species plasmid exchange between P.b, P.s, and C.p was observed, despite the fact that *C. perfringens* and P.b frequently co-existed under clinical circumstance (Supplementary Figure S4). Unlike P.s plasmids, none of the P.b plasmids coded for lethal toxins, whereas they play important roles in the virulence, adaptation, and metabolism of P.b. Particularly, group_a plasmids, which are present in one-third of P.b strains, carry virulence genes, as well as several organic metabolism and biosynthesis genes (Figure 6A). For instance, *corC* coded functions for magnesium and cobalt efflux (Huang et al., 2021); *tlyC* coded functions for phagosomal escape (Whitworth et al., 2005); *asrABC-nirC* cluster coded functions for microbial sulfidogenesis, which produces genotoxic hydrogen sulfide as a potential trigger of colorectal cancer (Czyzewski and Wang, 2012; Wolf et al., 2022). Of note, the rarely appearing group_f plasmids in P.b encoded two types of mosquito toxins (i.e., the Cry16Aa/Cry17Aa pair, and the novel neurotoxin PMP1) that are carried only by two strains isolated from Malaysia and Brazil, which are geographically distant but populated with mosquito species, indicating a gene acquisition with insect–bacteria coevolution (Contreras et al., 2019). A similar pattern of plasmid coevolution within species was observed for group_a plasmids, which were mostly carried by strains isolated from human feces in different geographic locations (Figure 6). Moreover, the Tn3 and IS1182 family transposases flanked in the toxin genes in group_f plasmids were sporadically presented in other plasmid groups distinct form group_f, providing strong evidence that the transposases in P.b contribute greatly in broadening of its gene pool for virulence and adaptation (Figure 6A). Up to date, only twenties P.b strains have had their genomes sequenced, and *in situ* clinical isolates are rare. Therefore, more efforts should be made in the future to isolate *in situ* clinical strains and to complete their whole genome sequences to obtain more accurate population genomic characteristics of the plasmids.

Conclusion

In conclusion, this study revealed that *P. bifermentans* carries a highly plastic pangenome as an emerging human pathogen. The core-genome of this species encodes at least three clostridial toxin homologs that are associated with extracellular matrix degradation and cell membrane damage, which can initiate extracellular infections. MGEs, particularly plasmids, have greatly contributed to the broadening of virulence, adaptation, and metabolism gene pools of this species. Although no lethally toxic strains in this species have been reported, the high prevalence of plasmids and transposons in the genomes highlights the potential risk for the emergence of lethal strains. Of note, this species may trigger colorectal cancer by producing genotoxic hydrogen sulfide in the gut tract. Inherent limitations arise in this study due to the absence of validation for specific clinically-related results through wet-bench experiments, and the scarcity of clinically-derived *P. bifermentans* genomes in public databases hampers extensive large-scale analyses. Nevertheless, this study significantly contributes to our understanding of the population genomics insights on phylogeny and pathogenesis associated with the emerging human pathogen *P. bifermentans*. The findings underscore the imperative for enhanced efforts in the isolation and sequencing of clinically-relevant *P. bifermentans* strains. Such endeavors are crucial to proactively mitigate the potential threats posed by these strains to human health.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

XC: Data curation, Funding acquisition, Writing – original draft. YP: Data curation, Writing – original draft. GY: Investigation, Methodology, Writing – review & editing. LF: Resources, Writing – review & editing. XT: Validation, Writing – review & editing. PH: Investigation, Writing – review & editing. YM: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. LX: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (41907214); Shenzhen Science and Technology Program (JCYJ20190808111610984); Natural Science Foundation of Guangdong Province (2022A1515011961); Special Program of Sustainable Development, Issued by the Science, Technology and Innovation Commission of Shenzhen Municipality (KCXFZ202002011006448).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1293206/full#supplementary-material

References

Aldape, M. J., Bayer, C. R., Bryant, A. E., and Stevens, D. L. (2016). A novel murine model of *Clostridium sordellii* myonecrosis: insights into the pathogenesis of disease. *Anaerobe* 38, 103–110. doi: 10.1016/j.anaerobe.2016.01.004

Barrett, L. F., Saragadam, S. D., DiMaria, C. N., and Delgado-Daza, A. (2020). Infection of a prosthetic knee joint with *Clostridium bifermentans*. *Oxf. Med. Case Rep.* 8, 256–258. doi: 10.1093/omcr/omaa057

Biswas, R., Raja, S., Sistla, S., Gopalakrishnan, M., and Saxena, S. K. (2018). Brain abscess and cervical lymphadenitis due to Paraclostridium bifermentans: a report of two cases. *Anaerobe* 51, 8–11. doi: 10.1016/j.anaerobe.2018.03.006

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Chen, P., Tao, L., Wang, T. Y., Zhang, J., He, A., Lam, K.-h., et al. (2018). Structural basis for recognition of frizzled proteins by *Clostridium difficile* toxin B. *Science* 360, 664–669. doi: 10.1126/science.aar1999

Chong, E., Winikoff, B., Charles, D., Agnew, K., Prentice, J. L., Limbago, B. M., et al. (2016). Vaginal and rectal *Clostridium sordellii* and *Clostridium perfringens* presence among women in the United States. *Obstet. Gynecol.* 127, 360–368. doi: 10.1097/AOG.0000000000001239

Contreras, E., Masuyer, G., Qureshi, N., Chawla, S., Dhillon, H. S., Lee, H. L., et al. (2019). A neurotoxin that specifically targets Anopheles mosquitoes. *Nat. Commun.* 10, 1–10. doi: 10.1038/s41467-019-10732-w

Couchman, E. C., Browne, H. P., Dunn, M., Lawley, T. D., Songer, J. G., Hall, V., et al. (2015). *Clostridium sordellii* genome analysis reveals plasmid localized toxin genes encoded within pathogenicity loci. *BMC Genomics* 16, 1–13. doi: 10.1186/s12864-015-1613-2

Czyzewski, B. K., and Wang, D. N. (2012). Identification and characterization of a bacterial hydrosulphide ion channel. *Nature* 483, 494–497. doi: 10.1038/nature10881

de Nies, L., Lopes, S., Busi, S. B., Galata, V., Heintz-Buschart, A., Laczny, C. C., et al. (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 9, 1–14. doi: 10.1186/s40168-020-00993-9

Edagiz, S., Lagace-Wiens, P., Embil, J., Karlowsky, J., and Walkty, A. (2015). Empyema caused by *Clostridium bifermentans*: a case report. *Can J Infect Dis Med* 26, 105–107. doi: 10.1155/2015/481076

Ezaki, T. (2009) *Family VII. Peptostreptococcaceae fam. nov. Bergey's Manual of Systematic Bacteriology 2nd*. New York: Springer:1008–1013.

Fang, H. T., Oberoi, A. S., He, Z. Q., Khanal, S. K., and Lu, H. (2021). Ciprofloxacin-degrading Paraclostridium sp. isolated from sulfate-reducing bacteria-enriched sludge: optimization and mechanism. *Water Res.* 191:116808. doi: 10.1016/j.watres.2021.116808

Ferrés, I., and Iraola, G. (2021). Protocol for post-processing of bacterial pangenome data using Pagoo pipeline. *STAR Protocols* 2:100802. doi: 10.1016/j.xpro.2021.100802

Geier, R. R., Rehberger, T. G., and Smith, A. H. (2021). Comparative genomics of *Clostridium perfringens* reveals patterns of host-associated phylogenetic clades and virulence factors. *Front. Microbiol.* 12:649953. doi: 10.3389/fmicb.2021.649953

Gonzalez-Astudillo, V., Asin-Ros, J., Moore, J., Uzal, F. A., and Navarro, M. A. (2022). *Paeniclostridium sordellii*–associated peripartum metritis in goats. *Vet. Pathol.* 60, 69–74. doi: 10.1177/03009858221133506

Hale, A., Kirby, J. E., and Albrecht, M. (2016). *Fatal spontaneous Clostridium bifermentans necrotizing endometritis: a case report and literature review of the pathogen*. Open Forum Infect Di. Oxford: Oxford University Press.

Hotze, E. M., Le, H. M., Sieber, J. R., Bruxvoort, C., McInerney, M. J., and Tweten, R. K. (2013). Identification and characterization of the first cholesterol-dependent cytolysins from gram-negative bacteria. *Infect. Immun.* 81, 216–225. doi: 10.1128/IAI.00927-12

Huang, Y. C., Jin, F., Funato, Y., Xu, Z. J., Zhu, W. L., Wang, J., et al. (2021). Structural basis for the Mg^{2+} recognition and regulation of the CorC Mg^{2+} transporter. *Sci. Adv.* 7:eabe6140. doi: 10.1126/sciadv.abe6140

Huang, Y. D., Wang, B., Yang, Y. G., Yang, S., Dong, M. J., and Xu, M. Y. (2022). Microbial carriers promote and guide pyrene migration in sediments. *J. Hazard. Mater.* 424:127188. doi: 10.1016/j.jhazmat.2021.127188

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 1–8. doi: 10.1038/s41467-018-07641-9

Jyothsna, T. S., Tushar, L., Sasikala, C., and Ramana, C. V. (2016). *Paraclostridium benzoelyticum* gen. nov., sp. nov., isolated from marine sediment and reclassification of *Clostridium bifermentans* as *Paraclostridium bifermentans* comb. nov. Proposal of a new genus *Paeniclostridium* gen. nov. to accommodate *Clostridium sordellii* and *Clostridium ghonii*. *Int. J. Syst. Evol. Micr.* 66, 1268–1274. doi: 10.1099/ijsem.0.000874

Klötzl, F., and Haubold, B. (2020). Phylonium: fast estimation of evolutionary distances from large samples of similar genomes. *Bioinformatics* 36, 2040–2046. doi: 10.1093/bioinformatics/btz903

Knight, D. R., Squire, M. M., Collins, D. A., and Riley, T. V. (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. *Front. Microbiol.* 7:2138. doi: 10.3389/fmicb.2016.02138

Kutsuna, R., Miyoshi-Akiyama, T., Mori, K., Hayashi, M., Tomida, J., Morita, Y., et al. (2019). Description of Paraclostridium bifermentans subsp. muricolitidis subsp. nov., emended description of Paraclostridium bifermentans (Sasi Jyothsna et al., 2016), and creation of Paraclostridium bifermentans subsp. bifermentans subsp. nov. *Microbiol. Immunol.* 63, 1–10. doi: 10.1111/1348-0421.12663

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. (2015). GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* 11:e1004941. doi: 10.1371/journal.pgen.1004941

Lee, H., Beilhartz, G. L., Kucharska, I., Raman, S., Cui, H., Lam, M. H. Y., et al. (2020). Recognition of semaphorin proteins by *P. Sordellii* lethal toxin reveals principles of receptor specificity in clostridial toxins. *Cells* 182, 345–356.e16. doi: 10.1016/j.cell.2020.06.005

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Ludueña, L. M., Anzuay, M. S., Angelini, J. G., Barros, G., Luna, M. F., Monge, M. P., et al. (2017). Role of bacterial pyrroloquinoline quinone in phosphate solubilizing ability and in plant growth promotion on strain Serratia sp. S119. *Symbiosis* 72, 31–43. doi: 10.1007/s13199-016-0434-7

Madden, J. C., Ruiz, N., and Caparon, M. (2001). Cytolysin-mediated translocation (CMT): a functional equivalent of type III secretion in gram-positive bacteria. *Cells* 104, 143–152. doi: 10.1016/S0092-8674(01)00198-2

Moore, R. J., and Lacey, J. A. (2019). Genomics of the pathogenic Clostridia. *Microbiol Spectr* 7:GPP3-0033-2018. doi: 10.1128/microbiolspec.GPP3-0033-2018

Neveling, O., Ncube, T. M., Ngxongo, Z. P., Chirwa, E. M., and Brink, H. G. (2022). Microbial precipitation of pb (ii) with wild strains of *Paraclostridium bifermentans* and *Klebsiella pneumoniae* isolated from an industrially obtained microbial consortium. *Int. J. Mol. Sci.* 23:12255. doi: 10.3390/ijms232012255

O'Brien, D. K., and Melville, S. B. (2004). Effects of Clostridium perfringens alpha-toxin (PLC) and perfringolysin O (PFO) on cytotoxicity to macrophages, on escape from the phagosomes of macrophages, and on persistence of *C. perfringens* in host tissues. *Infect. Immun.* 72, 5204–5215. doi: 10.1128/IAI.72.9.5204-5215.2004

Orrell, K. E., and Melnyk, R. A. (2021). Large clostridial toxins: mechanisms and roles in disease. *Microbiol. Mol. Biol. Rev.* 85, e00064–e00021. doi: 10.1128/MMBR.00064-21

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

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

Popoff, M. R. (2016). Toxins of histotoxic clostridia: *Clostridium chauvoei*, *Clostridium septicum*, *Clostridium novyi*, and *Clostridium sordellii*. *Clostridial Diseases of Animals*. 18, 21–43. doi: 10.1002/9781118728291.ch4

Popoff, M. R., and Bouvet, P. (2009). Clostridial toxins. *Future Microbiol.* 4, 1021–1064. doi: 10.2217/fmb.09.72

Portnoy, D. A., Jacks, P. S., and Hinrichs, D. J. (1988). Role of hemolysin for the intracellular growth of *Listeria monocytogenes*. *J. Exp. Med.* 167, 1459–1471. doi: 10.1084/jem.167.4.1459

Revitt-Mills, S. A., Vidor, C. J., Watts, T. D., Lyras, D., Rood, J. I., and Adams, V. (2019). Virulence plasmids of the pathogenic Clostridia. *Microbiol. Spectr.* 7:7. doi: 10.1128/microbiolspec.GPP3-0034-2018

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006

Vidor, C. J., Watts, T. D., Adams, V., Bulach, D., Couchman, E., Rood, J. I., et al. (2018). Clostridium sordellii pathogenicity locus plasmid pCS1-1 encodes a novel clostridial conjugation locus. *MBio* 9, e01761–e01717. doi: 10.1128/mBio.01761-17

Weinberg, M., and Séguin, P. (1918) *La gangrene gazeuse, Monographies de ITnstitut Pasteur*, Paris: Masson & Cie.

Weiss, W. J., Lenoy, E., Murphy, T., Tardio, L., Burgio, P., Projan, S. J., et al. (2004). Effect of srtA and srtB gene expression on the virulence of Staphylococcus aureus in animal models of infection. *J. Antimicrob. Chemoth.* 53, 480–486. doi: 10.1093/jac/dkh078

Whitworth, T., Popov, V. L., Yu, X.-J., Walker, D. H., and Bouyer, D. H. (2005). Expression of the *Rickettsia prowazekii* pld or tlyC gene in *Salmonella enterica* serovar typhimurium mediates phagosomal escape. *Infect. Immun.* 73, 6668–6673. doi: 10.1128/IAI.73.10.6668-6673.2005

Wolf, P. G., Cowley, E. S., Breister, A., Matatov, S., Lucio, L., Polak, P., et al. (2022). Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer. *Microbiome* 10, 1–16. doi: 10.1186/s40168-022-01242-x

Wouters, J. A., Hain, T., Darji, A., Hüfner, E., Wemekamp-Kamphuis, H., Chakraborty, T., et al. (2005). Identification and characterization of Di-and tripeptide transporter DtpT of *Listeria monocytogenes* EGD-e. *Appl. Environ. Microb.* 71, 5771–5778. doi: 10.1128/AEM.71.10.5771-5778.2005

Zerrouki, H., Rebiahi, S.-A., Elhabiri, Y., Fatmi, A., Baron, S. A., Pagnier, I., et al. (2021). Prevalence and antimicrobial resistance of Paeniclostridium sordellii in hospital settings. *Antibiotics* 11:38. doi: 10.3390/antibiotics11010038

Zhao, H. L., Wang, J. Q., Peng, Y., Cai, X. C., Liu, Y. D., Huang, W. Q., et al. (2022). Genomic insights from Paraclostridium bifermentans HD0315_2: general features and pathogenic potential. *Front. Microbiol.* 13:928153. doi: 10.3389/fmicb.2022.928153

TYPE Original Research
PUBLISHED 15 November 2023
DOI 10.3389/fmicb.2023.1272173

OPEN ACCESS

EDITED BY
Tulio de Lima Campos,
Oswaldo Cruz Foundation (Fiocruz), Brazil

REVIEWED BY
Carla Rodrigues,
Institut Pasteur, France
Jing Yang,
National Institute for Communicable Disease Control and Prevention (China CDC), China

*CORRESPONDENCE
Li Feng
✉ lifeng@zafu.edu.cn

RECEIVED 03 August 2023
ACCEPTED 01 November 2023
PUBLISHED 15 November 2023

CITATION
Feng L, Zhang M and Fan Z (2023) Population genomic analysis of clinical ST15 *Klebsiella pneumoniae* strains in China.
Front. Microbiol. 14:1272173.
doi: 10.3389/fmicb.2023.1272173

Population genomic analysis of clinical ST15 *Klebsiella pneumoniae* strains in China

Li Feng*, Mingcheng Zhang and Zhiyi Fan

Jiyang College, Zhejiang A&F University, Zhuji, China

ST15 *Klebsiella pneumoniae* (Kpn) is a growing public health concern in China and worldwide, yet its genomic and evolutionary dynamics in this region remain poorly understood. This study comprehensively elucidates the population genomics of ST15 Kpn in China by analyzing 287 publicly available genomes. The proportion of the genomes increased sharply from 2012 to 2021, and 92.3% of them were collected from the Yangtze River Delta (YRD) region of eastern China. Carbapenemase genes, including OXA-232, KPC-2, and NDM, were detected in 91.6% of the studied genomes, and 69.2% of which were multidrug resistant (MDR) and hypervirulent (hv). Phylogenetic analysis revealed four clades, C1 (KL112, 59.2%), C2 (mainly KL19, 30.7%), C3 (KL48, 0.7%) and C4 (KL24, 9.4%). C1 appeared in 2007 and was OXA-232-producing and hv; C2 and C4 appeared between 2005 and 2007, and both were KPC-2-producing but with different levels of virulence. Transmission clustering detected 86.1% (n = 247) of the enrolled strains were grouped into 55 clusters (2–159 strains) and C1 was more transmissible than others. Plasmid profiling revealed 88 plasmid clusters (PCs) that were highly heterogeneous both between and within clades. 60.2% (*n* = 53) of the PCs carrying AMR genes and 7 of which also harbored VFs. KPC-2, NDM and OXA-232 were distributed across 14, 4 and 1 PCs, respectively. The MDR-hv strains all carried one of two homologous PCs encoding *iucABCD* and *rmpA2* genes. Pangenome analysis revealed two major coinciding accessory components predominantly located on plasmids. One component, associated with KPC-2, encompassed 15 additional AMR genes, while the other, linked to OXA-232, involved seven more AMR genes. This study provides essential insights into the genomic evolution of the high-risk ST15 CP-Kpn strains in China and warrants rigorous monitoring.

KEYWORDS
***Klebsiella pneumoniae*, carbapenemase, molecular epidemiology, genomics, phylogenetics, pan-genome**

1. Introduction

Klebsiella pneumoniae (Kpn) causes a range of infections, including pulmonary, urinary tract, bloodstream, and surgical site infections (Paczosa and Mecsas, 2016). Carbapenems are commonly employed for treating severe infections caused by multidrug-resistant (MDR) *Enterobacteriaceae*, including *AmpC* β-lactamases and extended-spectrum β-lactamases (ESBLs). Unfortunately, the extensive use of carbapenems in recent years has expedited the emergence of resistant strains (Shrivastava et al., 2018).

As in many other countries, Kpn is a notifiable disease in China (Zhang et al., 2017; Kazmierczak et al., 2021; Lee et al., 2022). According to the China Antimicrobial Surveillance Network (CHINET) results, although the resistance rate of Kpn to carbapenems showed a steady downward trend from 2018 to 2021, the detection rate was still over 23% (Hu et al., 2018). The detection rate reached more than 30% in some areas, and the trend is slowly increasing. Previous epidemiology studies have shown that KPC-2 is the widest disseminated carbapenemase in China, and the dominant ST is ST11 (Zhang et al., 2017). However, ST15 Kpn becomes an emerging high-risk clone with frequent hospital outbreaks (Cienfuegos-Gallet et al., 2022). A multi-center study showed a shift in the dominant sequence type of carbapenemase-producing Kpn (CP-Kpn) bloodstream infections from ST11 to ST15 in northeast China, especially after the COVID-19 pandemic (Chen J. et al., 2021). ST15 Kpn has been reported to contain the *bla*$_{\text{OXA-232}}$ gene situated within the ColKP3-type (also known as ColE-type) plasmid in China (Yin et al., 2017; Shu et al., 2019; Chen Y. et al., 2021). Long-term nosocomial surveillance of OXA-48-like carbapenemases report in Zhejiang province, southeast China, from 2018 to 2021 showed that ST15 CP-Kpn isolates are the primary carriers in recent years (Zhang et al., 2023).

In addition to threats from CP-Kpn, infections due to hypervirulent Kpn have steadily increased over the last three decades (Russo and Marr, 2019). The hypervirulent strains are usually isolated from community-acquired infections and may cause a liver abscess, bloodstream infection, or meningitis, among other pathological conditions (Choby et al., 2020). The reported best-characterized virulence factors with experimental support for conveying the hypervirulent phenotype, including *iuc*, *iro*, *rmpA*, and *rmpA2*, are encoded by genes on hypervirulent (hv) plasmids (Zhu J. et al., 2021). Increasing occurrence of multidrug resistance (MDR) and hv Kpn (MDR-hvKpn) convergent clones is being observed(Wyres et al., 2020b). A public health concern is that virulence gene carriage has been reported to be 34.2% for CP-Kpn in China (Zhang et al., 2020).

Recent advancements in the whole-genome sequencing and extended applications of bioinformatic tools facilitate gathering information on thousands of bacterial species on their virulence factors (VF), antimicrobial resistance (AMR), and genetic relationship (Schürch et al., 2018). Comparative genomics of microbial genomes assists in understanding the genomic variations, the basis of diverse phenotypes (Wyres et al., 2020a). Although there has been a recent phylogenomics study on ST15 Kpn strains worldwide, only 9 strains were isolated from China (Rodrigues et al., 2023).

To delve into the genomic landscape of ST15 Kpn population in China, this study collected 287 genomes of clinical ST15 Kpn of China origin from the PATRIC database (Davis et al., 2020). Comparative genomic analyzes were performed to understand the spread of the ST15 Kpn strains across the country over 10 years. Then phylogenetic relationship, evolution, recent transmission, antimicrobial resistance and virulence gene profiling, pan-genome association, and plasmid content were screened.

2. Materials and methods

2.1. Genome collection and quality control

We retrieved all publicly available ST15 Kpn genome assemblies present in the PATRIC database on September 15, 2022 using the search terms "mlst = 15," "genome status = WGS," "host common name = Human," "isolation country = China" and "genome quality = good." These strains were all collected in China from 2012 to 2022. The corresponding metadata of the genomes was acquired from the PATRIC database and manually checked based on the NCBI Genbank database. Assembled genome and quality summary statistics were calculated with QUAST v5.2.0 and fastANI v1.33, respectively (Gurevich et al., 2013; Jain et al., 2018). All genomes passed the quality control with more than 99.5% of ANI and 85% of genome fraction with Kpn WSD411 (RefSeq: GCF_009884415.1) as an ST15 reference genome (Chen Y. et al., 2021). The MLST sequence type of each genome was also confirmed with Kleborate v2.2.0 (Lam et al., 2021).

2.2. Genome annotation

Genome annotation was performed using Prokka v1.13.4 (Seemann, 2014). Kleborate was used to identify the species identity, Kpn integrative conjugative element (ICEKp)-associated and plasmid-associated VF and AMR genes (Lam et al., 2021). Kleborate also assigns a virulence score and a resistance score for each genome. MOB-suite v3.0.3 was used to reconstruct plasmid content from each draft genome (Robertson and Nash, 2018). MOB-recon was used to analyze plasmid sequences, which includes MOB_typer to perform relaxase and replicon typing of plasmids, as well as generate MOB-cluster codes and host range information. The 28-bp fusion site was identified using the matchPattern function in the Biostrings R package with the specific sequence 'AGATCCGNAANNNNNNNN TTNCGGATCT' (Xu et al., 2021).

2.3. Phylogenetic and population structure analyzes

The core genome multi-alignment and SNP calling (cgSNP) was performed with Parsnp v1.2 from HarvestTools kit for the 287 genomes with collection dates and WSD411 as reference (Treangen et al., 2014). Pairwise SNP distances were calculated with SNP-sites v2.5.1 (Page et al., 2016). Phylogenetic trees were then constructed with RAxML v8.2.9 using the core genome SNP alignment after removing predicted recombination sites by Gubbins v2.1.0 (Stamatakis, 2014; Croucher et al., 2015). A general-time reversible nucleotide substitution model with a GAMMA correction for site variation was used for tree construction (bootstrap 1,000 with Lewis ascertainment correction). The output from Gubbins was loaded directly to BactDating v1.0.6, which accounts for branch-specific recombination rates, rather than simply ignoring recombinant regions (Didelot et al., 2014). Root-to-tip regression with a simultaneous inference of the best root location ($R^2 = 0.39$) and tip-date-randomization performed within BactDating demonstrated a temporal signal in the data. 100 million Markov chain Monte Carlo (MCMC) steps were performed to generate a time-resolved tree using the mixed model for clock rate.

Phylogenetic clades were identified using fastBAPS, and a core-genome SNP (cgSNP) threshold of 16 was selected to define the putative transmission relationship, respectively (Tonkin-Hill et al., 2019; Zhang et al., 2022). Furthermore, a cgMLST allele calling was performed using chewBBACA suite with a public 2,358-gene cgMLST scheme for *K. pneumoniae*/*variicola*/*quasipneumoniae* (Silva et al., 2018).[1] Ancestral sequence reconstruction of each internal node of the phylogenetic tree was performed using the R package phangorn (Schliep, 2011). Terminal branch lengths were the number of substitutions mapped to each terminal branch. Both phylogenetic tree and metadata were visualized with R package ggtree (Yu et al., 2018). A median-joining haplotype network was reconstructed by PopArt v1.7 (Leigh and Bryant, 2015).

2.4. Pangenome construction and coincident analysis

A pangenome was generated from all genomes using Panaroo with default parameters, resulting in a gene absence-presence matrix (Tonkin-Hill et al., 2020). Pangenome sequences retrieved by Panaroo were annotated with eggnog-mapper v2.1.2 using eggnogDB v5.0.2 (Huerta-Cepas et al., 2019). The antibiotic resistance and virulence factor genes were screened using the Comprehensive Antibiotic Resistance Database (CARD) and the virulence factor database (VFDB), utilizing a protein identity threshold of 80% (Liu et al., 2022; Alcock et al., 2023). The absence-presence matrix of the accessory genome was plotted by the Uniform Manifold Approximation and Projection (UMAP) algorithm with R package umap (McInnes et al., 2018).

To determine if antibiotic resistance genes are co-circulating with each other accessory gene and each other, we adopted the program Coinfinder v1.2.0 (Whelan et al., 2020). Briefly, Coinfinder detects genes that associate or dissociate with other genes using a Bonferroni-corrected Binomial exact test statistic of the expected and observed rates of gene–gene association. We first ran Coinfinder on our combined dataset to identify all coincident associated gene pairs. Then we reran Coinfinder using the query flag to look specifically at simultaneously associated gene pairs involving KPC-2, OXA-232, and NDM-1, respectively. Gephi v0.9.4 was used to visualize a coincident gene network with the Fruchterman-Reingold layout algorithm (Bastian et al., 2009).

3. Results

3.1. Overview of sequenced ST15 Kpn clinical isolates in China

This study obtained 287 ST15 Kpn genome assemblies from China after database screening on September 15, 2022 (Supplementary Table S1). Zhejiang province has the most samples, accounting for 56.5% (162/287), followed by Shanghai city samples, accounting for 27.2% (78/287). Notably, the Yangtze River Delta (YRD) region in eastern China, encompassing Zhejiang, Jiangsu, and Anhui provinces and Shanghai city, was the top region infected with ST15 Kpn, accounting for 92.3% (265/287) of all samples (Figures 1A,B). The distribution of ST15 Kpn samples by year showed a rapidly growing tendency, except during the COVID-19 pandemic, which dominated 2020 (Figure 1C). Despite that, 72.1% (207/287) of the ST15 Kpn samples were isolated between 2019 and 2021.

3.2. Phylogenetic and genomic characteristics

All ST15 Kpn strains were closely related with a maximum pairwise SNP distance of 222 SNPs, raising the possibility of clonal expansion of a common strain. The most-recent common ancestor of the 287 strains with isolation dates was estimated to emerge in August 2000 (95% confidence interval, April 1996 to October 2003) (Figure 2). The population of ST15 Kpn in China was further divided into four monophyletic clades based on a cgSNP/fastBAPS analysis that were C1, C2, C3, and C4 from top to bottom on the tree. There were 170 (59.2%) strains in C1, 88 (30.7%) in C2, 2 (0.7%) in C3, and 27 (9.4%) in C4, respectively. We inferred that C2 emerged first in May 2005, while the other three clusters emerged in the same year, 2007. Moreover, the latest sampling times for C1 and C2 are January 12, 2022, and December 1, 2021, respectively. In contrast, C3 and C4 have no new isolates after 2020 and 2019, respectively.

The capsule type (KL), VF and AMR characteristics revealed by Kleborate were further mapped on the phylogeny (Figure 2). A total of 11 distinct KL types were identified, with KL112 (59.2%, 170/287), KL19 (24.7%, 71/287), and KL24 (10.5%, 30/287) emerging as the predominant ones (Supplementary Figure S1). Significantly, unique capsule types, namely KL112, KL48, and KL24, corresponded to C1, C3, and C4, respectively. C2 displayed a diversity of nine KL types, with KL19 as the predominant one, accounting for 80.7% (71/88), and included three KL24 strains.

C1 and C4 displayed significantly higher virulence than C2 and C3 due to their aerobactin and yersiniabactin VFs (Fisher's exact test $p < 0.001$). Among the studied strains, 91.6% (263/287) were CP-Kpn, of which 33.1% (88/263) producing KPC-2, 57.6% (167/263) producing OXA-232, 1.1% (3/263) producing NDM, 1.5% (4/263) producing both KPC-2 and NDM, and 0.4% (1/263) producing both OXA-232 and NDM. Besides, 64.1% (184/287) of the dataset presented *iucABCD* and *rmpA2* genes, and from this 98.9% (182/184) were also carbapenemase producer. These MDR-hvKpn events included 156 C1, 2 C2 and 24 C4 strains. Associations were observed between carbapenemases and clades, with significant enrichment of OXA-232 in C1 and KPC-2 in both C2 and C4 (Fisher's exact test $p < 0.001$). Drug resistance also significantly varied among different clades (Fisher's exact test $p < 0.001$). We also found four strains DD02162 (C2), DD02172 (C4), K210279 (C1), and SCKP-LL83 (C2) exhibited the highest resistance scores of 3 for both carbapenemases and colistin resistance. Colistin resistance mechanisms included the presence of the colistin-resistant *mcr-1* gene in SCKP-LL83 and inactivated mutations in the *mgrB* gene in the other three strains.

To assess whether the genotypes of ST15 Kpn strains in China differed from those of strains isolated from other global areas, we further included the genomic data of 293 strains collected from other parts of the world (Supplementary Table S2) (Rodrigues et al., 2023). Similar to the strains in China, the most common KL type in these global strains was KL112 (49.1%, 144/293), but the proportion of KL24 (35.2%, 103/293) was higher than KL19 (5.1%, 15/293). The maximum-likelihood phylogenetic tree of all 580 strains showed that C1-C4 in China all had highly homologous strains from other global regions, especially Asia and Europe (Supplementary Figure S2). Except for few strains in C1 and C2, most of the strains collected from China were monophyletic in the phylogenetic tree.

[1] https://www.cgmlst.org/ncs/schema/2187931/

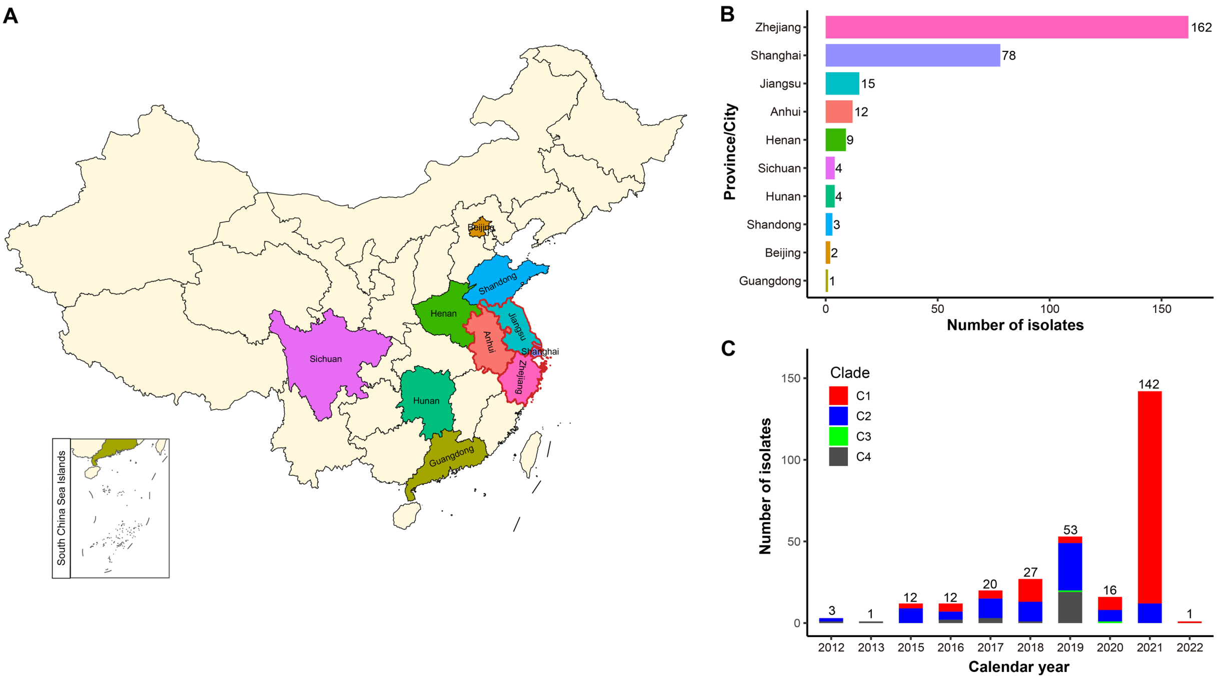


FIGURE 1
Geographic and temporal distribution of the studied ST15 Kpn strains in China. **(A)** The map of China shows the sampling areas for the enrolled strains, with the boundaries of the Yangtze River Delta (YRD) region highlighted in red. **(B)** Bar plot shows the prevalence of ST15 Kpn strains in 10 provinces and cities of China. **(C)** The stacked bar chart shows the frequency of ST15 strains sequenced in China each year from 2012 to January 2022 and the proportion of the four clades.

3.3. Clonal transmission evidence in China's ST15 Kpn isolates

The diversity metrics of their subtrees were calculated to assess the difference in transmissibility and capability of causing active disease in infected hosts among the four clades. The C1 phylogeny had significantly shorter terminal branch lengths than the other clades (Figure 3A, Wilcoxon test $p < 0.001$). Strains belonging to C1 were genetically more similar than those belonging to different clades, as indicated by the smaller median pairwise SNP distance (Figure 3B, Wilcoxon test $p < 0.001$). Furthermore, we explored the distribution of potential transmission clusters using a range spanning 1 ~ 100 SNPs of maximum pairwise SNP distance thresholds to define a transmission cluster (Figure 3C). Notably, the proportion of strains belonging to transmission clusters was significantly higher among C1 strains than in other clades.

The genomic distance matrices computed on the cgMLST concatenate and cgSNP alignment was significantly correlated (Mantel test, $p < 0.001$; Spearman test R = 0.81, $p < 0.001$). According to a recently published molecular epidemiology study of CP-Kpn in Shanghai, the clonal clusters were defined using a cutoff of 16 SNPs (Zhang et al., 2022). Here, 245 (85.4%) of the 287 strains were detected in 18 clonal clusters, ranging in size from two to 159 (Supplementary Figure S3). Furthermore, a cgMLST typing analysis retrieved a similar clustering result that 247 (86.7%) of the 287 strains into 15 clonal clusters at an allele distance threshold of ten, which has been used to correctly group all surgical intensive care unit outbreak strains in a hospital in Beijing (Zhou et al., 2017).

The cgSNP clustering rate of C1 (97.6%, 166/170) was significantly higher than that of C2 (64.8%, 57/88), C3 (0, 0/2), and C4 (81.5%, 22/27) (chi-square test, $p < 0.001$). In addition, 77.8% (14/18) of clonal clusters involved transmissions that occurred within 1 year. The longest for the other four clonal clusters is 6.2 years, followed by 3.0, 1.9, and 1.7 years. Notably, there were five cgSNP-based clonal clusters including 167 strains involving recent transmissions across different provinces (Figure 3D). Four clusters (cluster 1, 7, 11 and 14) were located in the YRD region, and only one cluster (cluster 14) was the transmission between Beijing and Zhejiang. The largest cluster (cluster 1), including 93.5% (159/170) strains of C1, isolated from 2015 to 2022, revealed a large-scale continuous spreading of the ST15 OXA-232-CP-Kpn of C1 in the YRD region.

3.4. Plasmids profiling based on the complete genomes

To reveal the plasmid communities shared among China's ST15 Kpn population, we adopted three tools in MOB-suite to all genomes: MOB-recon for plasmid sequence identification, MOB-typer for plasmid typing, MOB-cluster for plasmid clustering, respectively. A total of 2,101 plasmids were detected and grouped into 88 plasmid clusters (PCs) at a mash distance threshold of 0.05 (Supplementary Table S3A). Only 31 of these PCs contained more than 5 plasmids, indicating the complexity of plasmid content. An average number of PCs per genome was 7.4 (between 1 and 12), and the number increased from C3 (average = 3) to C2 (average = 3.6) to C1 (average = 9.7) to C4 (average = 5.0), with a statistically significant

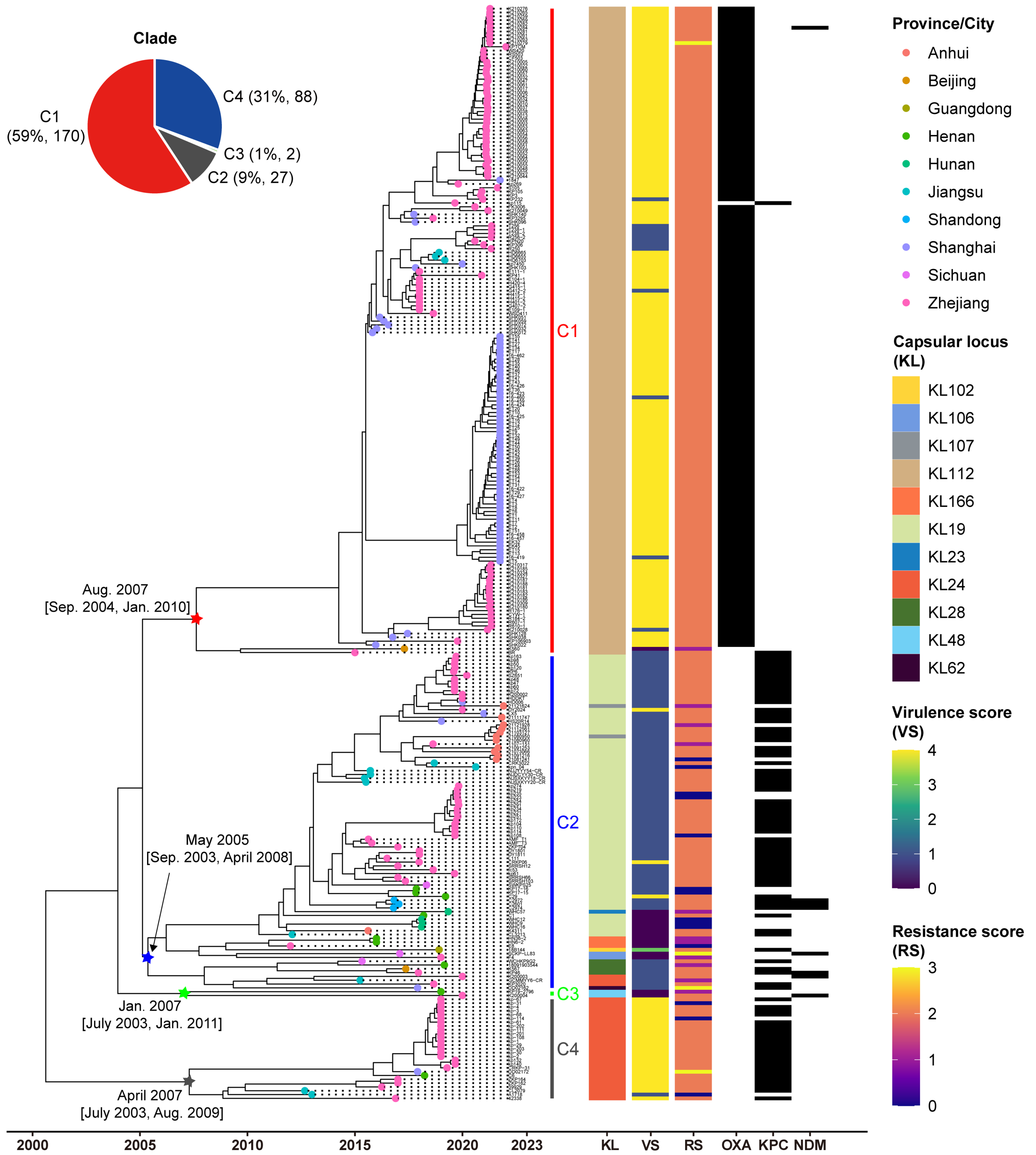


FIGURE 2
Low genetic diversity in China's ST15 Kpn population. Dated phylogenomic tree of the core SNP analysis for China's ST15 Kpn strains. The tree was constructed using BactDating and corrected for recombination using Gubbins. Individual nodes were colored by provinces as defined in the legend. The distinct clades, capsular locus (KL) type, virulence scores, and resistance scores of strains identified by Klebroate, presence of carbapenemase genes (OXA-232, KPC-2, and NDM) were shown on the tree (from inner to outer strips). The virulence score is based on the presence of *ybt*, *clb*, and *iuc* as follows: 0, none present; 1, *ybt* only; 2, *clb* without *iuc* (regardless of *ybt*; however, *ybt* is almost always present when *clb* is); 3, *iuc* only; 4, *iuc* and *ybt* without *clb*; and 5, all three genes present. Resistance scores are calculated as follows: 0 = no ESBL or carbapenemase, 1 = ESBL without carbapenemase (regardless of colistin resistance); 2 = carbapenemase without colistin resistance (regardless of ESBL); 3 = carbapenemase with colistin resistance (regardless of ESBL). The estimated origin times and 95% CI of four clades are shown at the relevant nodes. The time scale is indicated at bottom.

difference (ANOVA test, $p < 0.001$) (Figure 4A). The hierarchical clustering based on the presence and absence of PCs among all strains showed a clear separation of the plasmid content between C1, C2 and C4, as well as between KPC-2 and OXA-232. Furthermore, since there was no common PC present in all strains, we focused on the core PC within each clade (CC-PC), that was, the PC that appear in more than 80% of the members. We found that C1 and C4 had 8 and 3 CC-PCs, respectively, while C2 and C3 had none. Based on the

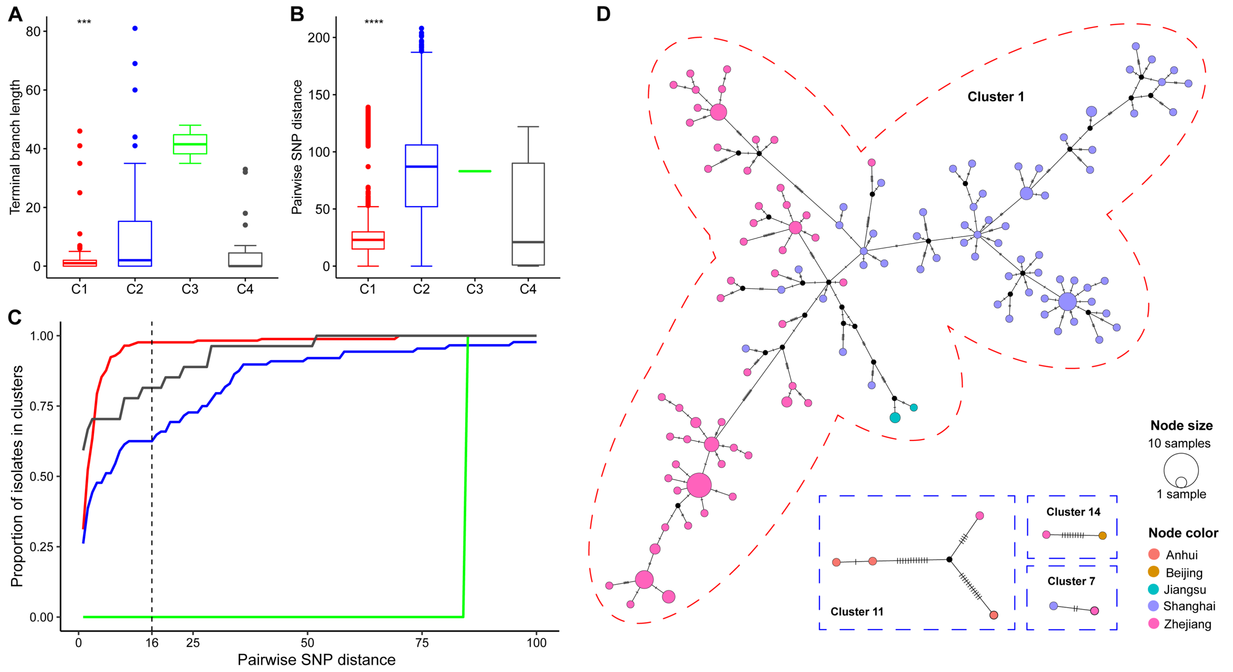


FIGURE 3
Genomic epidemiology of China's ST15 Kpn. **(A)** The distribution of terminal branch lengths for different clades (colored as in Figure 2). **(B)** The distribution of pairwise SNP distances for different clades. Three and four asterisks indicate Wilcoxon test *p*-values smaller than 0.001 and 1e-04, respectively. **(C)** Proportion of isolates from each clade that belong to clusters (y axis) defined at different thresholds for maximum pairwise SNP distances (x axis). **(D)** Median-Joining network generated for four cross-regional transmission clusters using PopArt. Clusters are outlined with a dashed circle representing the color of the clade in which their strain belongs. A circle represents a haplotype. The circle area represents the frequency of strains, and hatch marks across branches indicate mutational steps on the edges. Black dots indicate inferred missing isolates. The color inside each circle represented the isolation area.

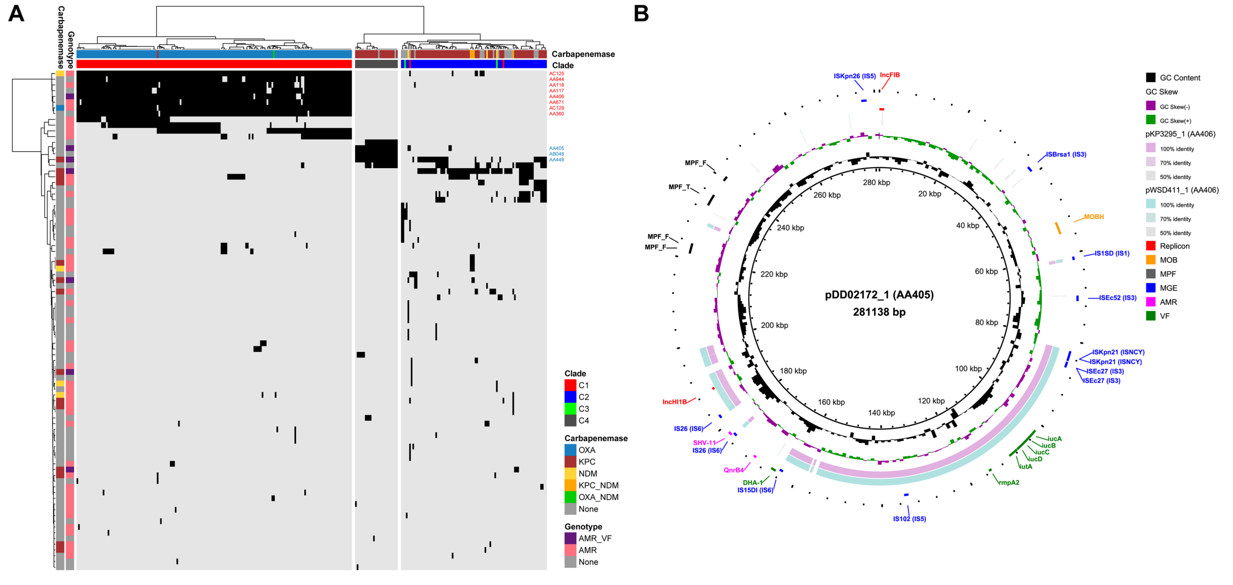


FIGURE 4
Plasmid profiling of China's ST15 Kpn. **(A)** Heat map of hierarchical clustering based on the presence (black) and absence (light gray) of 88 PCs (row) among 287 strains (column). The colors on the top of the heat map represent the clade to which the strain belongs and the carbapenemase it encodes. The colors on the left side of the heat map indicate whether each PC encodes AMR and/or VF genes and the carbapenemase it encodes. The 8 CC-PCs in C1 and 3 CC-PCs in C4 are marked. **(B)** Alignment of the complete plasmids of AA405 and AA406 using BLAST Ring Image Generator (BRIG) (Alikhan et al., 2011). The plasmid pDD02172_1 in AA405 is used as the reference. The outer colored labels refer to the annotation of replicon, MOB, mating pair formation (MPF), mobile genetic element (MGE), AMR and VF gene, respectively.

16 complete genomes, we found that 4 CC-PCs belong to 3 Inc. replicon families (FIB, HI1B and U), 4 CC-PCs belonged to 2 Col replicon types. Besides, 2 CC-PCs were conjugative, 4 CC-PCs were mobilizable, and 5 CC-PCs were non-mobilizable.

We identified 60.2% (53/88) of the PCs carrying AMR genes, with an average of 2.8 AMR genes per PC (ranging from 1 to 17) (Supplementary Table S3B). KPC-2 was found in 14 PCs including a CC-PC (AA448, IncU-type), 12 of which were conjugative or mobilizable. NDM was present in four PCs including a CC-PC (AC125, IncFIB-type), two of which were conjugative or mobilizable, while OXA-232 was exclusively detected in a mobilizable CC-PC (AC129, rep_cluster_1195). Moreover, we found the colistin resistance gene *mcr-1* in two PCs within C2: AA378, which carried one AMR gene, and AA738, which encoded 17 AMR genes. Notably, seven of these AMR PCs also carried VF genes (Supplementary Table S3C). Among them, the VF genes *iucABCD* and *rmpA2* coexisted on 183 plasmids, forming two F-type PCs: AA405 in C4 (KPC-2-producing) and AA406 in C1 (OXA-232-producing). Sequence alignment of two representative plasmids, pDD02172_1 (AA405) and pWSD411_1 (AA406), revealed their homology with a mean identity of 86.9% and coverage of 44.3% (pDD02172_1 as the reference) (Figure 4B). There were IS sequences belonging to the IS*NCY*, IS*3* and IS*6* families at both ends of the homologous region containing the VFs. However, AA405 was conjugative, while AA406 lacked both relaxase and mate-pair formation, making it non-mobilizable. Interestingly, we detected the non-mobilizable pWSD411_1 has the potential to co-transfer with a conjugative F-type plasmid pWSD411_2 for both sharing the 28-bp fusion site (Xu et al., 2021).

3.5. An open structure of China's ST15 Kpn pan-genome

To characterize the genomic diversity of the analyzed 287 ST15 Kpn genomes, a pan-genome was constructed. This pan-genome consists of 4,539 core and 4,377 accessory genes. The simulated gene accumulation curves showed that the numbers of the core genes decreased continually with the addition of new strains, as expected when sampling more diverging genomes of a species (Supplementary Figure S4). Heap's law modeling ($n = \kappa N^{\gamma}$) of the gene presence-absence revealed a γ value of 0.1 less than 1, demonstrating the open state of the pan-genome. Displaying the genomes using a UMAP approach directly on the absence-presence of accessory genes showed a clear separation of the three main clades (C1, C2, and C4) identified by fastBAPS based on the cgSNP (Figure 5A). Although C2 is closer to C1 in terms of genetic distance, the accessory genome composition of C2 is more comparable to that of C4. This phenomenon suggested that the composition of accessory genes between different strain clusters may be related to the mechanisms of carbapenem resistance.

3.6. Coincident genes associated with carbapenem-resistance genes

To further explore the coincident gene relationships within the pan-genome, a gene co-occurrence network was inferred by Coinfinder (Figure 5B). It contained a total of 228,491 significant gene-to-gene relationships, including 1,691 coincident genes, accounting for 38.6% (1,691/4377) of all accessory genes (Bonferroni-corrected binomial exact test, $p < 0.05$). These associated gene pairs were further clustered into 25 associate components. There were two large components in the network, containing 1,037 and 508 genes, respectively, and occupied 91.4% (1,545/1691) of all coincident genes. The first component included the KPC-2 with other 15 AMR genes [$bla_{CTX\text{-}M\text{-}15}$, *catB11*, *tet(A)*, *dfrA12*, *aph(3″)-Ia*, *sul1*, *aadA22*, *qacEdelta1*, $bla_{OXA\text{-}1}$, *msrE*, *mphE*, *armA*, *qnrB4*, $bla_{DHA\text{-}1}$ and *aac(3)-IId*] contributed to the resistance to cephalosporin, penam, tetracycline, diaminopyrimidine, aminoglycoside, sulfonamide, phenicol, streptogramin, macrolide, fluoroquinolone and cephamycin antibiotics. The second component included the OXA-232 with other seven AMR genes [*TEM-1*, *sul2*, *aph(6)-Id*, *aph(3″)-Ib*, *arr-2*, *qnrB2* and *rmtF*] contributed to the resistance to penem, cephalosporin, monobactam, sulfonamide, aminoglycoside, rifamycin and fluoroquinolone antibiotics. All other components were far smaller, ranging from 2 to 42 genes. Two other gene clusters also included antibiotic resistance genes; one of 9 genes included *qnrS1*, and one of 42 had *ramR*.

When we only examined coincident gene–gene relationships involving the three carbapenemases, including KPC-2, OXA-232, and NDM-1, we identified 383, 391, and 0 coincident genes, respectively. Six genes directly coincident with KPC-2 were AMR genes: *sul1*, *mphE*, *DHA-1*, *qnrB4*, *armA,* and *msrE*. The genes directly coincident with OXA-232 included three AMR genes (*rmtF*, *qnrB2,* and *arr-2*), and two VF genes (*iucA* and *rmpA*), related to aerobactin and mucoid phenotype A regulation, respectively. In addition, we found that all the AMR and VF genes coincident with OXA-232 and KPC-2 were located on the plasmid according to the complete genomes WSD411, DD02162, and KP46 (Supplementary Tables S2B,C).

By comparing the functions between the KPC-2's and OXA-232's coincident genes, we found that genes related to post-translational modification, protein turnover, and chaperones (COG category code O) and intracellular trafficking, secretion, and vesicular transport (COG category code U) were much more prevalent in the KPC-2's coincident genes (Supplementary Figure S5A). In contrast, genes related to replication, recombination, and repair (COG category code L), inorganic ion transport and metabolism (COG category code P), and signal transduction mechanisms (COG category code T) were more likely to be included in the OXA-232's coincident genes. Furthermore, 71.5% (274/383) of the KPC-2's coincident genes and 80.1% (313/391) of the OXA-232's coincident genes were plasmid-mediated (Supplementary Figure S5B). There was no significant difference in the distribution of KPC-2's and OXA-232's coincident genes on chromosome and plasmid (chi-square test, $p = 0.402$).

4. Discussion

In this study, by screening all public ST15 Kpn genomes, we found that the isolation frequency of ST15 Kpn in China has continued to increase over the past decade. The ST15 Kpn in China originated in 2000 and has differentiated into four distinct clades. The origin of these clades was as early as 2005 and as late as 2008, and some strains are still emerging in 2022. The predominant KL types in the studied strains are KL112, KL19, and KL24. A comparison with strains from other global regions indicates similarities in KL types,

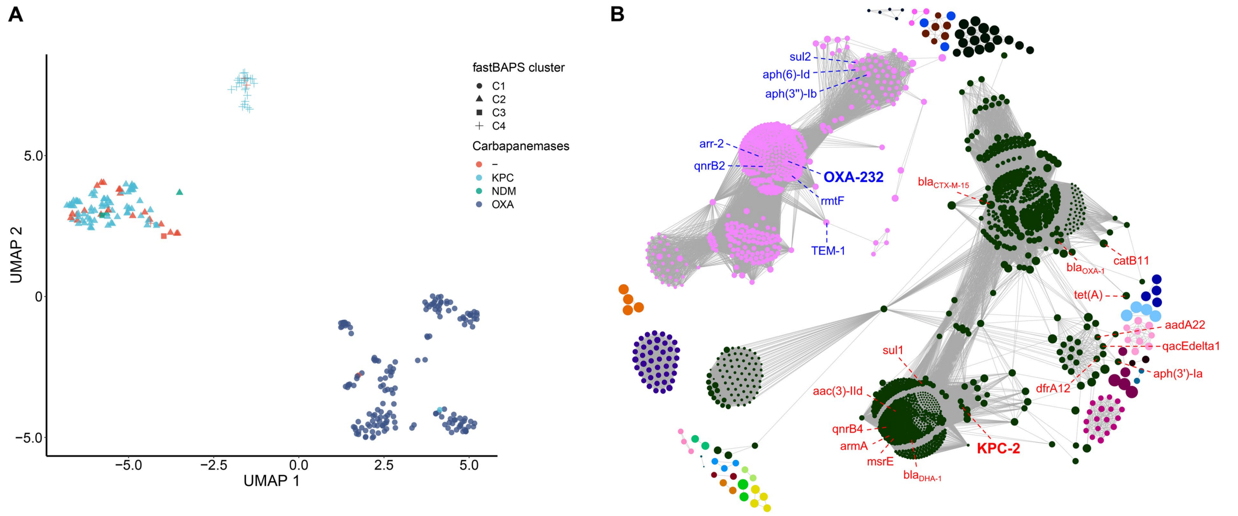


FIGURE 5
Pan-genome modeling of China's ST15 Kpn population. **(A)** Visualization of the UMAP two-dimensional representation of the pan-genome. Strains in different clades were shown with different shapes and colored by different carbapenemase genes. **(B)** Network diagram created with Gephi using output from Coinfinder carried on China's ST15 Kpn pan-genome. Nodes are colored by connected components (coincident gene sets). The size of a node is proportional to the gene's D value (Whelan et al., 2020).

with KL112 being the most common (Rodrigues et al., 2023). This suggests a global distribution of certain KL types within ST15 Kpn population. Notably, most new cases occurred in the YRD region, eastern China, between 2019 and 2021. With the high resolution provided by genomics, we revealed that up to 85% of isolates were due to recent transmission. C1 and C4 displayed higher virulence, likely contributing to the severity of infections they cause. Recently, numerous nosocomial outbreaks of ST15 Kpn were reported in the YRD region including Hangzhou, Lishui, Wenzhou, Yancheng, Jiaxing and Shanghai (Li et al., 2019; Jia et al., 2021; Zhu Z. et al., 2021; Huang et al., 2022; Wu et al., 2023; Zhang et al., 2023). The YRD region is one of the most economically active regions in China and attracts a large number of migrant workers from Yunnan, Sichuan, and Anhui provinces every year. This finding suggested that there was a high level of transmission of ST15 CP-Kpn between hospitals by patient transfer. However, the transmission of CP-Kpn within and between hospitals remains largely unexplored in China (Cienfuegos-Gallet et al., 2022). Meanwhile, it should be pointed out that the YRD region has more medical resources than other central and western regions in China, which may be one of the reasons why most of the currently sequenced strains originate from this region. Accordingly, we speculate that ST15 Kpn has been widely disseminated in China in recent years.

The emergence and expansion of CP-Kpn have resulted in a bottleneck in effective antimicrobial treatment (Zong et al., 2021). Worryingly, through AMR gene testing, we found that 92% of the enrolled ST15 Kpn strains are CP-Kpn, and 69% of which are MDR-hvKpn with *iuc* and *rmpA2*. MDR and hv are typically observed in separate Kpn populations. However, convergent strains with both properties have been documented and potentially pose a high risk to public health in the form of invasive infections with limited treatment options (Arcari and Carattoli, 2023). OXA-232 is the most detected carbapenemase in China's ST15 CP-Kpn, followed by KPC-2, while NDM is relatively rare. Notably, the co-occurrence of NDM with both KPC-2 and OXA-232 has already appeared. Therefore, the emergence of NDM has become a growing public health threat and represents a new challenge for the treatment of infectious diseases (Gao et al., 2020).

Notably, our pan-genome analysis provides valuable insights into the relationship between genomic diversity, clade-specific differentiation, and the presence of carbapenem resistance genes among ST15 Kpn. First, phylogenetic analysis based on the cgSNP showed the emergence of distinct clades (C1, C2, C3, and C4) is associated with the presence of these carbapenemase genes. Second, the observed open pan-genome structure reflects the remarkable diversity within ST15 Kpn and indicates that they can exchange genetic material (Holt et al., 2015). The distinct clades exhibit varying accessory gene profiles, and most of the accessory genes are coincides with KPC-2 or OXA-232 and located on the plasmids. Therefore, the pan-genome of China's ST15 CP-Kpn has already differentiated into KPC-2-type and OXA-232-type structures at both core and accessory genomes. In addition, these coincident accessory genes include some AMR genes confer to aminoglycoside, sulfonamide, cephalosporin and fluoroquinolone, and aerobactin and regulation VF genes. Indeed, fluoroquinolone resistance appears to confer a fitness advantage to high-risk clones of various species, particularly among the elderly and individuals with prolonged healthcare center exposure, which is a known risk factor for acquiring additional antibiotic resistance genes (Redgrave et al., 2014; Fuzi et al., 2020). These evidence suggested an adaptive evolution of plasmid-mediated large-scale horizontal gene transfer among China's ST15 Kpn strains.

We found 88 different PCs in China's ST15 Kpn strains and high variation in plasmid content among different clades. C1 and C4 display a more stable and clade-specific plasmid repertoire with a higher number of CC-PCs. In contrast, C2 and C3 lack CC-PCs, indicating a less stable plasmid composition. The overrepresentation of F and Col plasmids and high heterogeneity of small plasmids in China's ST15 Kpn was also similar to a recent global ST15 Kpn research (Rodrigues et al., 2023). Notably, up to 60% of the PCs in our study encoded at least one AMR gene, with a maximum of 17 AMR genes, and 7 of them also carried VFs. Comparing KPC-2 and NDM, there is only 1 PC carrying OXA-232, suggesting that the spread of the KPC-2 and NDM is more complex than that of OXA-232 in China's ST15 Kpn (Zhang et al., 2023). There is one mobilizable CC-PC encoding OXA-232 and one conjugative CC-PC encoding KPC-2 in C1 and C4, respectively.

We emphasize that the *iucABCD* and *rmpA2* genes in all MDR-hvKpn genomes are located on plasmids. Although these plasmids belong to two PCs, one was conjugative and the other could co-transfer with a conjugative F-type plasmid in the same genome (Xu et al., 2021). They have a certain degree of homology and might be formed through recombination mediated by IS sequences (Acman et al., 2022). The presence of both VF and AMR genes, especially the carbapenemase genes, and *iuc* and *rmpA2* VFs on plasmids enables simultaneous transfer in a single event and potentially rapid emergence of MDR-hvKpn clone (Tang et al., 2020). Moreover, the presence of colistin resistance genes on specific plasmids in C2 strains is a concerning development, as colistin is often considered a last-resort antibiotic (Zong et al., 2021).

We acknowledge the imperfect nature of the ST15 Kpn dataset we used. First, only the PATRIC database was used for sample screening, which is largely biased and commonly not well-curated. There was insufficient diversity among China's ST15 strains included in the study. Second, the available metadata can significantly impact the dating estimation and may not be correct. Third, there was no related experiment to demonstrate both drug resistance and virulence potential from genomic detection.

In conclusion, this study provides a comprehensive view of the molecular epidemiology and genetic diversity in the China's ST15 Kpn population. Our findings demonstrated that clonal transmission was the leading cause of the increasing incidence of infections due to the ST15 CP-Kpn during the past 5 years. The variety of the cgSNP-based phylogeny, the composition of accessory genes, and the plasmid profiles correlated to the two different carbapenem genes, OXA-232 and KPC-2. These findings provide essential perspectives into ST15 CP-Kpn and highlight the urgent need for medical institutions to strengthen surveillance to prevent these novel strains from further disseminating in hospital settings and the community.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

LF: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. MZ: Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. ZF: Data curation, Formal analysis, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Jiyang College of Zhejiang A&F University under Grant [number RQ1911F12].

Acknowledgments

The authors thank the authors and laboratories who submitted sequences to the PATRIC Database.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1272173/full#supplementary-material

SUPPLEMENTARY TABLE S1
Metadata and genotyping results of 287 public China's ST15 Kpn genomes from the PATRIC database.

SUPPLEMENTARY TABLE S2
Metadata of 293 public ST15 Kpn genomes worldwide.

SUPPLEMENTARY TABLE S3
Plasmid-typing **(A)** results from the MOB-suite software and AMR **(B)** and VF **(C)** genes content of each plasmid.

SUPPLEMENTARY FIGURE S1
Statistics of all serotypes.

SUPPLEMENTARY FIGURE S2
Phylogenetic tree of 287 China's and 293 global ST15 Kpn strains.

SUPPLEMENTARY FIGURE S3
Statistics of members of all transmission clusters based on a threshold of 16 SNPs.

SUPPLEMENTARY FIGURE S4
Simulations of the increase of the pan-genome size and the decrease of core-genome size.

SUPPLEMENTARY FIGURE S5
COG category annotation **(A)** and genomic localization **(B)** statistics of the KPC-2's and OXA-232's coincident gene.

References

Acman, M., Wang, R., van Dorp, L., Shaw, L. P., Wang, Q., Luhmann, N., et al. (2022). Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene *bla*$_{\mathrm{NDM}}$. *Nat. Commun.* 13:1131. doi: 10.1038/s41467-022-28819-2

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., et al. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 51, D690–D699. doi: 10.1093/nar/gkac920

Alikhan, N.-F., Petty, N. K., Zakour, N. L. B., and Beatson, S. A. (2011). BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Arcari, G., and Carattoli, A. (2023). Global spread and evolutionary convergence of multidrug-resistant and hypervirulent *Klebsiella pneumoniae* high-risk clones. *Pathog. Glob. Health* 117, 328–341. doi: 10.1080/20477724.2022.2121362

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *AJS* 3, 361–362. doi: 10.1609/icwsm.v3i1.13937

Chen, Y., Fang, L., Yang, Y., Yan, R., Fu, Y., Shen, P., et al. (2021). Emergence of carbapenem-resistant *Klebsiella pneumoniae* harbouring Bla OXA-48-like genes in China. *J. Med. Microbiol.* 70:001306. doi: 10.1099/jmm.0.001306

Chen, J., Hu, C., Wang, R., Li, F., Sun, G., Yang, M., et al. (2021). Shift in the dominant sequence type of carbapenem-resistant *Klebsiella pneumoniae* bloodstream Infection from ST11 to ST15 at a medical Center in Northeast China, 2015–2020. *IDR* 14, 1855–1863. doi: 10.2147/IDR.S311968

Choby, J., Howard-Anderson, J., and Weiss, D. (2020). Hypervirulent *Klebsiella pneumoniae*–clinical and molecular perspectives. *J. Intern. Med.* 287, 283–300. doi: 10.1111/joim.13007

Cienfuegos-Gallet, A. V., Zhou, Y., Ai, W., Kreiswirth, B. N., Yu, F., and Chen, L. (2022). Multicenter genomic analysis of carbapenem-resistant *Klebsiella pneumoniae* from Bacteremia in China. *Microbiol. Spectr.* 10:e0229021. doi: 10.1128/spectrum.02290-21

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Res.* 43:e15. doi: 10.1093/nar/gku1196

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., et al. (2020). The PATRIC bioinformatics resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* 48, D606–D612. doi: 10.1093/nar/gkz943

Didelot, X., Gardy, J., and Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* 31, 1869–1879. doi: 10.1093/molbev/msu121

Fuzi, M., Rodriguez Baño, J., and Toth, A. (2020). Global evolution of pathogenic bacteria with extensive use of fluoroquinolone agents. *Front. Microbiol.* 11:271. doi: 10.3389/fmicb.2020.00271

Gao, H., Liu, Y., Wang, R., Wang, Q., Jin, L., and Wang, H. (2020). The transferability and evolution of NDM-1 and KPC-2 co-producing *Klebsiella pneumoniae* from clinical settings. *EBioMedicine* 51:102599. doi: 10.1016/j.ebiom.2019.102599

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

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., et al. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl. Acad. Sci.* 112, E3574–E3581. doi: 10.1073/pnas.1501049112

Hu, F., Zhu, D., Wang, F., and Wang, M. (2018). Current status and trends of antibacterial resistance in China. *Clin. Infect. Dis.* 67, S128–S134. doi: 10.1093/cid/ciy657

Huang, J., Chen, X., Yang, J., Zhao, Y., Shi, Y., Ding, H., et al. (2022). Outbreak of KPC-producing *Klebsiella pneumoniae* ST15 strains in a Chinese tertiary hospital: resistance and virulence analyses. *J. Med. Microbiol.* 71:001494. doi: 10.1099/jmm.0.001494

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9

Jia, H., Zhang, Y., Ye, J., Xu, W., Xu, Y., Zeng, W., et al. (2021). Outbreak of multidrug-resistant OXA-232-producing ST15 *Klebsiella pneumoniae* in a teaching Hospital in Wenzhou, China. *IDR* 14, 4395–4407. doi: 10.2147/IDR.S329563

Kazmierczak, K. M., Karlowsky, J. A., de Jonge, B. L., Stone, G. G., and Sahm, D. F. (2021). Epidemiology of carbapenem resistance determinants identified in meropenem-nonsusceptible *Enterobacterales* collected as part of a global surveillance program, 2012 to 2017. *Antimicrob. Agents Chemother.* 65, 10–1128. doi: 10.1128/AAC.02000-20

Lam, M. M. C., Wick, R. R., Watts, S. C., Cerdeira, L. T., Wyres, K. L., and Holt, K. E. (2021). A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat. Commun.* 12:4188. doi: 10.1038/s41467-021-24448-3

Lee, Y.-L., Ko, W.-C., and Hsueh, P.-R. (2022). Geographic patterns of global isolates of carbapenem-resistant *Klebsiella pneumoniae* and the activity of ceftazidime/avibactam, meropenem/vaborbactam, and comparators against these isolates: results from the antimicrobial testing leadership and surveillance (ATLAS) program, 2020. *Int. J. Antimicrob. Agents* 60:106679. doi: 10.1016/j.ijantimicag.2022.106679

Leigh, J. W., and Bryant, D. (2015). PopART: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Li, X., Ma, W., Qin, Q., Liu, S., Ye, L., Yang, J., et al. (2019). Nosocomial spread of OXA-232-producing *Klebsiella pneumoniae* ST15 in a teaching hospital, Shanghai, China. *BMC Microbiol.* 19:235. doi: 10.1186/s12866-019-1609-1

Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50, D912–D917. doi: 10.1093/nar/gkab1107

McInnes, L., Healy, J., and Melville, J. (2018). *Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426*.

Paczosa, M. K., and Mecsas, J. (2016). *Klebsiella pneumoniae*: going on the offense with a strong defense. *Microbiol. Mol. Biol. Rev.* 80, 629–661. doi: 10.1128/MMBR.00078-15

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial. Genomics.* 2:e000056. doi: 10.1099/mgen.0.000056

Redgrave, L. S., Sutton, S. B., Webber, M. A., and Piddock, L. J. (2014). Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends Microbiol.* 22, 438–445. doi: 10.1016/j.tim.2014.04.007

Robertson, J., and Nash, J. H. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial. Genomics* 4:e000206. doi: 10.1099/mgen.0.000206

Rodrigues, C., Lanza, V. F., Peixe, L., Coque, T. M., and Novais, Â. (2023). Phylogenomics of globally spread clonal groups 14 and 15 of *Klebsiella pneumoniae*. *Microbiol. Spectr.* 11:e0339522. doi: 10.1128/spectrum.03395-22

Russo, T. A., and Marr, C. M. (2019). Hypervirulent *klebsiella pneumoniae*. *Clin. Microbiol. Rev.* 32, 10–1128. doi: 10.1128/CMR.00001-19

Schliep, K. P. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi: 10.1093/bioinformatics/btq706

Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L., and Goering, R. V. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin. Microbiol. Infect.* 24, 350–354. doi: 10.1016/j.cmi.2017.12.016

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

Shrivastava, S. R., Shrivastava, P. S., and Ramasamy, J. (2018). World health organization releases global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *J. Med. Soc.* 32:76. doi: 10.4103/jms.jms_25_17

Shu, L., Dong, N., Lu, J., Zheng, Z., Hu, J., Zeng, W., et al. (2019). Emergence of OXA-232 Carbapenemase-producing *Klebsiella pneumoniae* that carries a pLVPK-like virulence plasmid among elderly patients in China. *Antimicrob. Agents Chemother.* 63, e02246–e02218. doi: 10.1128/AAC.02246-18

Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., et al. (2018). chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. *Microbial. Genomics* 4:e000166. doi: 10.1099/mgen.0.000166

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Tang, M., Kong, X., Hao, J., and Liu, J. (2020). Epidemiological characteristics and formation mechanisms of multidrug-resistant hypervirulent *Klebsiella pneumoniae*. *Front. Microbiol.* 11:581543. doi: 10.3389/fmicb.2020.581543

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W., and Corander, J. (2019). Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* 47, 5539–5549. doi: 10.1093/nar/gkz361

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* 21:180. doi: 10.1186/s13059-020-02090-4

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

Whelan, F. J., Rusilowicz, M., and McInerney, J. O. (2020). Coinfinder: detecting significant associations and dissociations in pangenomes. *Microbial. Genomics* 6:e000338. doi: 10.1099/mgen.0.000338

Wu, X., Li, X., Yu, J., Shen, M., Fan, C., Lu, Y., et al. (2023). Outbreak of OXA-232-producing carbapenem-resistant *Klebsiella pneumoniae* ST15 in a Chinese teaching hospital: a molecular epidemiological study. *Frontiers in cellular and infection. Microbiology* 13:1229284. doi: 10.3389/fcimb.2023.1229284

Wyres, K. L., Lam, M. M. C., and Holt, K. E. (2020a). Population genomics of *Klebsiella pneumoniae*. *Nat. Rev. Microbiol.* 18, 344–359. doi: 10.1038/s41579-019-0315-1

Wyres, K. L., Nguyen, T. N., Lam, M., Judd, L. M., van Vinh Chau, N., Dance, D. A., et al. (2020b). Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from south and Southeast Asia. *Genome Med.* 12, 1–16. doi: 10.1186/s13073-019-0706-y

Xu, Y., Zhang, J., Wang, M., Liu, M., Liu, G., Qu, H., et al. (2021). Mobilization of the nonconjugative virulence plasmid from hypervirulent *Klebsiella pneumoniae*. *Genome Med.* 13:119. doi: 10.1186/s13073-021-00936-5

Yin, D., Dong, D., Li, K., Zhang, L., Liang, J., Yang, Y., et al. (2017). Clonal dissemination of OXA-232 Carbapenemase-producing *Klebsiella pneumoniae* in neonates. *Antimicrob. Agents Chemother.* 61, e00385–e00317. doi: 10.1128/AAC.00385-17

Yu, G., Lam, T. T.-Y., Zhu, H., and Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* 35, 3041–3043. doi: 10.1093/molbev/msy194

Zhang, Y., Chen, C., Wu, J., Jin, J., Xu, T., Zhou, Y., et al. (2022). Sequence-based genomic analysis reveals transmission of antibiotic resistance and virulence among Carbapenemase-producing *Klebsiella pneumoniae* strains. *mSphere* 7, e00143–e00122. doi: 10.1128/msphere.00143-22

Zhang, Y., Jin, L., Ouyang, P., Wang, Q., Wang, R., Wang, J., et al. (2020). Evolution of hypervirulence in carbapenem-resistant *Klebsiella pneumoniae* in China: a multicentre, molecular epidemiological analysis. *J. Antimicrob. Chemother.* 75, 327–336. doi: 10.1093/jac/dkz446

Zhang, R., Liu, L., Zhou, H., Chan, E. W., Li, J., Fang, Y., et al. (2017). Nationwide surveillance of clinical carbapenem-resistant *Enterobacteriaceae* (CRE) strains in China. *EBioMedicine* 19, 98–106. doi: 10.1016/j.ebiom.2017.04.032

Zhang, Y., Yang, X., Liu, C., Huang, L., Shu, L., Sun, Q., et al. (2023). Increased clonal dissemination of OXA-232-producing ST15 *Klebsiella pneumoniae* in Zhejiang, China from 2018 to 2021. *Infect. Dis. Poverty* 12:25. doi: 10.1186/s40249-023-01051-w

Zhou, H., Liu, W., Qin, T., Liu, C., and Ren, H. (2017). Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Klebsiella pneumoniae*. *Front. Microbiol.* 8:371. doi: 10.3389/fmicb.2017.00371

Zhu, Z., Huang, H., Xu, Y., Wang, M., Lv, J., Xu, L., et al. (2021). Emergence and genomics of OXA-232-producing *Klebsiella pneumoniae* in a hospital in Yancheng, China. *J. Glob. Antimicrob. Resist.* 26, 194–198. doi: 10.1016/j.jgar.2021.05.015

Zhu, J., Wang, T., Chen, L., and Du, H. (2021). Virulence factors in hypervirulent *Klebsiella pneumoniae*. *Front. Microbiol.* 12:642484. doi: 10.3389/fmicb.2021.642484

Zong, Z., Feng, Y., and McNally, A. (2021). Carbapenem and colistin resistance in *Enterobacter*: determinants and clones. *Trends Microbiol.* 29, 473–476. doi: 10.1016/j.tim.2020.12.009

TYPE Original Research
PUBLISHED 23 November 2023
DOI 10.3389/fmicb.2023.1292897

Check for updates

OPEN ACCESS

EDITED BY
Tulio de Lima Campos,
Oswaldo Cruz Foundation (Fiocruz), Brazil

REVIEWED BY
Nagatoshi Fujiwara,
Tezukayama University, Japan
Sanjay Gautam,
International Vaccine Institute, Republic of Korea

*CORRESPONDENCE
Demin Cao
✉ caodm@ymun.edu.cn
Xidai Long
✉ sjtulongxd@ymun.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 12 September 2023
ACCEPTED 01 November 2023
PUBLISHED 23 November 2023

CITATION
Zhu X, Lu Q, Li Y, Long Q, Zhang X, Long X and Cao D (2023) Contraction and expansion dynamics: deciphering genomic underpinnings of growth rate and pathogenicity in *Mycobacterium*.
Front. Microbiol. 14:1292897.
doi: 10.3389/fmicb.2023.1292897

Contraction and expansion dynamics: deciphering genomic underpinnings of growth rate and pathogenicity in *Mycobacterium*

Xiaoying Zhu[1,2†], Qunfeng Lu[3,4†], Yulei Li[1], Qinqin Long[1], Xinyu Zhang[1], Xidai Long[1,2*] and Demin Cao[1*]

[1]Clinical Pathological Diagnosis & Research Center, The Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, Guangxi, China, [2]Medical College, Guangxi University, Nanning, Guangxi, China, [3]Modern Industrial College of Biomedicine and Great Health, Youjiang Medical University for Nationalities, Baise, Guangxi, China, [4]School of Medical Laboratory Sciences, Youjiang Medical University for Nationalities, Baise, Guangxi, China

Background: *Mycobacterium* bacteria, encompassing both slow growth (SGM) and rapid growth mycobacteria (RGM), along with true pathogenic (TP), opportunistic pathogenic (OP), and non-pathogenic (NP) types, exhibit diverse phenotypes. Yet, the genetic underpinnings of these variations remain elusive.

Methods: Here, We conducted a comprehensive comparative genomics study involving 53 *Mycobacterium* species to unveil the genomic drivers behind growth rate and pathogenicity disparities.

Results: Our core/pan-genome analysis highlighted 1,307 shared gene families, revealing an open pan-genome structure. A phylogenetic tree highlighted clear boundaries between SGM and RGM, as well as TP and other species. Gene family contraction emerged as the primary alteration associated with growth and pathogenicity transitions. Specifically, ABC transporters for amino acids and inorganic ions, along with quorum sensing genes, exhibited significant contractions in SGM species, potentially influencing their distinct traits. Conversely, TP strains displayed contraction in lipid and secondary metabolite biosynthesis and metabolism-related genes. Across the 53 species, we identified 26 core and 64 accessory virulence factors. Remarkably, TP and OP strains stood out for their expanded mycobactin biosynthesis and type VII secretion system gene families, pivotal for their pathogenicity.

Conclusion: Our findings underscore the importance of gene family contraction in nucleic acids, ions, and substance metabolism for host adaptation, while emphasizing the significance of virulence gene family expansion, including type VII secretion systems and mycobactin biosynthesis, in driving mycobacterial pathogenicity.

KEYWORDS

***Mycobacterium*, pathogenicity, growth rate, contraction/expansion, virulence factors, evolution**

1 Introduction

Within the phylum Actinomycetota, the genus *Mycobacterium* is diverse, comprising over 190 species. The cell walls of the *Mycobacterium* genus are characterized by a lipid-rich outer layer containing high concentrations of mycolic acid (Dulberger et al., 2020). Their unique composition necessitates the use of acid-fast staining to emphasize their resistance to acidic conditions, distinguishing them from other cell types (Pennington et al., 2021). However, bacteria within this genus exhibit highly distinctive phenotypic traits at various levels. Various types of mycobacteria are widely distributed across diverse environments, such as soil, water bodies, fish, amphibians, mammals, primates, and humans. Their classification into rapid grower mycobacteria (RGM) and slow grower mycobacteria (SGM) is determined by the growth of visible colonies on solid culture media within 7 days (Philley and Griffith, 2015). In addition to the well-known pathogens *M. tuberculosis* (causing tuberculosis) and *M. leprae* (causing leprosy), there are other species such as *M. marinum* and *M. shottsii* that induce diseases in fish and amphibians (Gauthier et al., 2021, 2022). Recent years have witnessed the emergence of opportunistic infections caused by diverse non-tuberculosis mycobacteria (NTM), introducing a new challenge in clinical therapy (Gopalaswamy et al., 2020; Ratnatunga et al., 2020). RGM are predominantly environmental saprophytes, whereas all pathogenic bacteria fall within the slow growth mycobacteria category. However, opportunistic pathogens are distributed across all types.

Distinct genetic factors underpin the pathogenicity, growth rate, and remarkable adaptability of bacteria to diverse habitats. Conducting comparative genomic analysis is an effective approach to comprehensively decipher the genotype–phenotype relationship in bacteria and understand the interplay between species' genomic evolution and their environment. Our previous study revealed that the capacity to utilize diverse carbohydrates, enabling *Listeria monocytogenes* and *L. ivanovii* to effectively adapt to and exploit various carbohydrate sources in their environment, represents a crucial genetic trait contributing to their pathogenicity (Lu et al., 2022). By conducting a genomic comparison between SGM and RGM, Bachmann et al. discovered that the absence of *livFGMH*, *shaAC-G*, and ABC transporter operons, which are associated with amino acid and ion transport in SGM, may contribute to alterations in their growth rate (Bachmann et al., 2019). Recent comparative genomic studies have highlighted the pivotal role of bacterial genome expansions or contractions in driving modifications in their biological capabilities (Baroncelli et al., 2016). Notably, the *pe* and *ppe* gene families, with the majority of their members secreted through the Type-VII secretion system, undergo significant expansion in pathogenic *Mycobacterium* species (Cole et al., 1998). These gene families play vital roles in maintaining iron homeostasis. For example, PE5-PPE4 is implicated in mycobactin-mediated iron acquisition, whereas PPE36 and PPE37 are linked to heme-iron acquisition, among other functions (Tufariello et al., 2016; Tullius et al., 2019). These expansions within the gene families are likely to contribute to the virulence and pathogenicity of *Mycobacterium*. Nonetheless, our current comprehension of how gene family expansion and contraction influence the evolution of pathogenicity and environmental adaptability in mycobacteria remains incomplete. Further exploration at a broader genomic level is warranted.

Recently, a considerable number of complete genomes from various *Mycobacterium* species have been sequenced. This wealth of genomic data provides a valuable opportunity to obtain comprehensive insights into the genetic foundation and evolutionary attributes underlying the intricate diversity of *Mycobacterium* phenotypes. Such insights hold potential implications for the prevention and treatment of *Mycobacterium* infections. In this study, we conducted a comparative genomic study using representing complete genomes of 53 *Mycobacterium* species. In the present study, we performed a comparative genomic analysis utilizing complete genomes representing 53 distinct *Mycobacterium* species. These species were stratified into two groups according to their growth rates: SGM and RGM. Furthermore, an additional categorization was established based on their pathogenicity, dividing them into three groups: Totally Pathogens (TP), Opportunity Pathogens (OP), and Non-Pathogens (NP). Subsequently, core/pan-genome analysis and functional annotation were conducted, yielding insights into the genomic structure and functional makeup of bacteria within this genus. Next, a phylogenetic tree was constructed utilizing single-copy core genes, and the gene families undergoing contraction/expansion for each node were estimated. Additional analysis, considering gene function and protein interactions, elucidates the role of gene expansion/contraction events in influencing pathogenicity and environmental adaptability. Lastly, we investigated the distribution and expansion/contraction of virulence factors within each mycobacterial species. Our findings uncover the pivotal role of gene family contraction in the transition from RGM to SGM, as well as in the acquisition of pathogenicity.

2 Materials and methods

2.1 Data retrieval and genome management

The complete-genome sequences of 53 *Mycobacterium* species were obtained from the genome database at NCBI,[1] with data retrieved until May 5, 2023. To ensure analytical consistency, we re-annotated the structure of each genome using the same pipeline. Specifically, the open reading frames of each genome were identified using Prodigal V2.6.3 (Hyatt et al., 2010). The identification of tRNAs was conducted with tRNAscan-SE 1.3.1, while rRNAs were identified using rnammer 1.2 (Lagesen et al., 2007). Key genome attributes, including genome size, GC-content, number of coding sequences (CDS), growth type, pathogenicity, and other pertinent features, were documented and are outlined in Table 1.

2.2 Core- and pan-genome analysis

The homologous genes were identified using OrthoFinder version 2.5.4, employing an all-*vs*-all BLASTp search based on the proteome of each genome (Emms and Kelly, 2019). Subsequently, we computed

[1] ftp://ftp.ncbi.nih.gov/genomes/

TABLE 1 The genome features of 53 representing *Mycobacterium* species included in this study.

Genus	Strain ID	Accession number	Genome size (bp)	GC content (%)	# of scaffold	# of plasmid	CDS	rRNA	tRNA	Growth type[a]	Pathogenicity[b]
M. tuberculosis	H37Rv	GCA 000195955.2	4,411,532	65.61	1	0	3,981	3	45	SGM	TP
M. avium	104	GCA 000014985.1	5,475,491	68.99	1	0	5,175	3	46	SGM	OP
M. fortuitum	JCM 6387	GCA 022179545.1	6,406,072	66.19	1	0	6,134	6	55	RGM	OP
M. intracellulare	ATCC 13950	GCA 000277125.1	5,402,402	68.1	1	0	5,003	3	47	SGM	OP
M. kansasii	ATCC 12478	GCA 000157895.2	6,577,228	66.23	1	1	5,723	3	46	SGM	TP
M. leprae	TN	GCA 000195855.1	3,268,203	57.8	1	0	3,896	3	45	SGM	TP
M. chelonae	CCUG 47445	GCA 001632805.1	5,029,817	63.92	1	0	4,827	3	47	RGM	OP
M. marinum	E11	GCA 000723425.2	6,450,522	65.74	1	1	5,328	6	48	SGM	TP
M. ulcerans	Agy99	GCA 000013925.2	5,805,761	65.39	1	1	5,397	3	45	SGM	TP
M. haemophilum	DSM 44634 ATCC 29548	GCA 000340435.3	4,235,765	63.95	1	0	3,897	3	45	RGM	OP
M. abscessus	GZ002	GCA 004028015.1	5,082,434	64.14	1	1	4,918	3	47	RGM	OP
M. gallinarum	JCM 6399	GCA 010726765.1	6,301,681	65.61	1	1	6,152	6	46	RGM	NP
M. branderi	JCM 12687	GCA 010728725.1	5,979,623	66.49	1	1	5,742	6	46	SGM	OP
M. conspicuum	JCM 14738	GCA 010730195.1	6,237,139	67.39	1	0	5,652	3	48	SGM	OP
M. heidelbergense	JCM 14842	GCA 010730745.1	5,050,576	67.94	1	0	4,583	3	45	SGM	OP
M. canettii	CIPT 140010059	GCA 000253375.1	4,482,059	65.62	1	0	3,960	3	45	SGM	TP
M. heckeshornense	JMUB5695	GCA 016861545.1	4,865,109	65.9	1	0	4,557	6	46	SGM	OP
M. frederiksbergense	LB 501 T	GCA 012223425.1	6,713,618	67.07	1	3	6,521	6	46	RGM	NP
M. kubicae	NJH MKUB1	GCA 014263315.1	5,463,479	66.28	1	2	5,075	3	49	SGM	OP
M. shottsii	JCM 12657	GCA 010728525.1	5,973,149	65.53	1	0	5,625	3	47	SGM	TP
M. lentiflavum	ATCC 51985	GCA 022374895.2	6,163,560	65.94	1	3	5,650	3	47	SGM	OP
M. lacus	JCM 15657	GCA 010731535.1	5,092,988	66.95	1	0	4,605	3	45	SGM	OP
M. parmense	JCM 14742	GCA 010730575.1	5,952,912	68.39	1	0	5,435	3	49	SGM	NP
M. saskatchewanense	JCM 13016	GCA 010729105.1	6,008,916	68.28	1	0	5,574	3	48	SGM	OP
M. liflandii	ASM001	GCA 022354805.1	6,167,296	65.57	1	0	5,641	3	47	SGM	TP
M. pseudoshottsii	JCM 15466	GCA 003584745.1	6,061,597	65.64	1	0	5,430	3	47	SGM	TP
M. florentinum	JCM 14740	GCA 010730355.1	6,219,859	66.38	1	0	5,805	3	47	SGM	OP
M. colombiense	CECT 3035	GCA 002105755.1	5,581,643	68.09	1	0	5,300	3	47	SGM	OP

(Continued)

TABLE 1 (Continued)

Genus	Strain ID	Accession number	Genome size (bp)	GC content (%)	# of scaffold	# of plasmid	CDS	rRNA	tRNA	Growth type[a]	Pathogenicity[b]
M. seoulense	JCM 16018	GCA 010731595.1	5,531,300	68.29	1	0	5,087	3	48	SGM	NP
M. shinjukuense	JCM 14233	GCA 010730055.1	4,504,020	67.79	1	0	3,984	3	45	SGM	OP
M. salmoniphilum	DSM 43276	GCA 004924335.1	4,776,697	64.29	1	0	4,572	3	56	RGM	NP
M. noviomagense	JCM 16367	GCA 010731635.1	4,779,798	65.76	1	0	4,504	6	45	SGM	OP
M. stomatepiae	JCM 17783	GCA 010731715.1	6,210,822	65.96	1	0	5,981	3	48	SGM	OP
M. lepromatosis	FJ924	GCA 000975265.2	3,271,694	57.89	1	0	3,806	3	46	SGM	TP
M. dioxanotrophicus	PH-06	GCA 002157835.1	8,080,416	66.46	1	4	7,725	9	83	RGM	NP
M. riyadhense	NTM	GCA 016864455.1	6,772,223	65.1	1	0	5,998	3	48	SGM	TP
M. mantenii	JCM 18113	GCA 010731775.1	6,185,541	66.88	1	0	5,674	3	48	SGM	OP
M. paraterrae	DSM 45127	GCA 022430545.2	5,522,624	65.55	1	0	5,193	6	48	SGM	NP
M. paraseoulense	JCM 16952	GCA 010731655.1	6,085,955	67.9	1	0	5,641	3	46	SGM	NP
M. marseillense	FLAC0026	GCA 002285715.1	5,285,642	67.84	1	1	4,842	3	47	SGM	OP
M. shigaense	JCM 32072	GCA 002356315.1	5,232,660	67.26	1	0	4,871	3	47	SGM	OP
M. litorale	NIIDNTM18	GCA 014218295.1	5,634,149	68.88	1	0	5,407	6	47	RGM	NP
M. spongiae	FSD4b-SM	GCA 018278905.1	5,581,157	65.56	1	0	4,472	3	45	SGM	NP
M. paraintracellulare	M011	GCA 016756055.1	5,357,612	68.12	1	1	4,968	3	46	SGM	NP
M. paragordonae	49,061	GCA 003614435.1	7,224,251	66.89	1	4	6,411	3	47	SGM	OP
M. stephanolepidis	NJB0901	GCA 002356335.1	4,994,485	63.95	1	0	4,916	3	49	RGM	NP
M. grossiae	DSM 104744	GCA 008329645.1	5,681,602	70.45	1	1	5,382	6	46	RGM	NP
M. saopaulense	EPM10906	GCA 001456355.1	4,649,175	64.8	1	0	4,488	3	49	RGM	NP
M. vicinigordonae	24	GCA 013466425.1	6,266,765	65.35	1	0	5,577	3	46	SGM	NP
M. basiliense	901,379	GCA 900292015.1	5,607,630	65.02	1	0	4,696	3	48	SGM	OP
M. novum	JCM 6391	GCA 010726505.1	4,458,926	68.58	1	0	4,161	9	46	SGM	NP
M. ostraviense	FDAARGOS 1613	GCA 021183725.1	6,127,734	66.27	1	0	5,536	3	46	SGM	NP
M. senriense	TY59	GCA 019668465.1	5,831,451	67.08	3	0	5,339	3	47	SGM	NP

[a] SGM, slow growth mycobacteria; RGM, rapid growth mycobacteria.

[b] TP, totally pathogen; OP, opportunity pathogen; NP, not pathogen.

the proportion of homologous genes both between species and within species using the following formulas:

$$\eta = \frac{2N_s}{\left(N_a + N_b\right)} \times 100\%, \tag{1}$$

where η is the proportion of homologous genes, N_s is the number of homologous genes, N_a is the number of proteins of one genome, and N_b is the number of proteins of another genome. The results were visualized with a heatmap by r-ggplot2.

Pan-genome size can be determinated by a prediction using Heap's law, which is formulated as:

$$\mathrm{y} = A_1 x^{-B1} + C_1, \tag{2}$$

where y is the pan-genome size, x is the number of genomes used, and *A1, B1, C1* are the fitting parameters.

Core-genome size can be determinated by a prediction using power law, which is formulated as:

$$\mathrm{y} = A_2 x^{B2} + C_2, \tag{3}$$

where y is the core-genome size, x is the number of genomes used, and *A2, B2, C2* are the fitting parameters. The fitted curve of core and pan-genome were displayed using R version 4.2.1.

2.3 Phylogenetic analysis

Multiple amino acid sequence alignments of 507 single-copy genes from the 53 *Mycobacterium* species were generated using MAFFT v7.490 (Katoh and Standley, 2013). Highly divergent sites were subsequently removed using Clipkit v1.3.0 (Steenwyk et al., 2020). Subsequently, a maximum likelihood (ML) phylogenetic tree was constructed using the best-fit substitution model JTT+F+R7, as determined by ModelFinder (Kalyaanamoorthy et al., 2017), implemented in IQ-TREE v1.6.12 (Nguyen et al., 2015). Divergence times between species were estimated using the MCMC tree program in paml-v4.10.6 (Yang, 2007). To deduce alterations in gene family size, we utilized the program CAFE5-5.1.0 (Mendes et al., 2020). The topology of the ML tree was visualized using ggtree v3.6.2 (Yu et al., 2016). The growth type and pathogenicity of each species were displayed on the right side of the tree, while gene family contraction/expansion events were depicted using pie charts on the corresponding nodes of the tree.

2.4 Functional annotation of core- and accessory-genome

To annotate the genes within the core and accessory genome using Clusters of Orthologous Groups (COG), a whole-genome BLASTp search was performed against both the NCBI COG database (v20) and EggNOG v5.0 (Huerta-Cepas et al., 2019), employing an E-value cutoff of 1e-5. Subsequently, the results were integrated using a Python script.

2.5 Annotation of virulence factor and antibiotic resistance genes

The proteins encoded by genes from the 53 mycobacterial species underwent a BLASTp search against the virulence factor database (VFDB) to identify potential virulence-related genes (Liu et al., 2022). For the identification of antibiotic resistance genes within the genomes of the 53 *Mycobacterium* species, the genes were aligned against the comprehensive antibiotic resistance database (CARD) (Alcock et al., 2019). The BLASTp search was conducted with parameters set to 90% identity, 90% coverage, and an E-value cutoff of 1e-5. The outcomes were visualized using ComplexHeatmap v2.14.0 (Gu, 2022).

2.6 Analysis of protein–protein interactions

Protein–protein interaction networks for the sets of genes were constructed using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING v12.0, https://string-db.org/, accessed on 9 July, 2023) (Szklarczyk et al., 2021). Additionally, Cytoscape 3.9.1 was employed for visualizing these networks (Shannon et al., 2003).

2.7 Data availability statement

All data analyzed and generated throughout this study have been comprehensively integrated into both the manuscript and Supplementary material.

3 Results and discussion

3.1 Genome statistics and general features

The *Mycobacterium* genus exhibits significant species diversity and notable variation in phenotypic characteristics. The List of Prokaryotic Names with Standing in Nomenclature database (LPSN, https://www.bacterio.net/genus/mycobacterium, accessed until June 27, 2023) documents 195 taxa of *Mycobacterium*. To comprehensively explore the genomic diversity of the *Mycobacterium* genus, this study meticulously selected and analyzed complete-level genomes from 53 *Mycobacterium* species (Table 1). The genomes within this genus featured elevated GC content, with an average of 63.17%. The genome of *L. leprae* had the lowest GC content (57.8%), while *M. grossiae* had the highest GC content (70.45%) among them. The presence of plasmids was infrequent within *Mycobacterium* genomes, with only fifteen out of fifty-three species detected with plasmids. All strains were classified into SGM and RGM based on their growth rates. Out of these *Mycobacterium* species, 41 were SGM and 12 were RGM. Furthermore, the 53 species were categorized into TP, OP and NP based on their virulence. Among these, eleven species were categorized as TP, encompassing *M. tuberculosis, M. canettii, M. kansasii, M. leprae, M. leproma, M. ulcerans*—organisms that infect humans or other mammals, and *M. marinum, M. shottsii, M. liflandii, M. pseudoshottsii*, which infect fish or amphibians (Gauthier et al., 2021; Bartlett et al., 2022). A total of twenty-four species were classified as OP, implicated in causing lung or other infections in immunocompromised patients. The remaining eighty species were NP species, which rarely cause infection disease. In the natural context, pathogenic bacteria or specific parasitic bacteria typically exhibit smaller genomes compared to free-living species (Ochman and Davalos, 2006). This principle can also be extended to the analysis of *Mycobacterium* genomes.

3.2 Composition and functional characteristics of the pan-genome in 53 *Mycobacterium* species

To assess the genetic diversity among *Mycobacterium* species, we conducted an analysis of homologous genes for each pair of species (Figure 1A, Supplementary Table S1). *M. tuberculosis* and *M. canettii* exhibited the highest proportion of homologous genes, reaching 97.27%. This observation signifies a close relationship between these two species (Supply et al., 2013). It is evident that *M. chelonae, M. stephanolepidis, M. salmoniphilum, M. saopaulense, M. abscessus* exhibit noteworthy distinctions from other species, yet they share a closer relationship among themselves. In addition, the smallest genomes, *M. leprae* and *M. lepromatosis,* exhibited the lowest proportion of homologous genes shared with other species. Regarding homology within the genome, *M. dioxanotrophicus* displayed 19.52% homologous genes, whereas *M. haemophilum* had 9.16% ones. Notably, pathogens such as *M. canettii* (9.61%), *M. tuberculosis* (9.62%), and *M. leprae* (9.33%) exhibited a low proportion of homologous genes within their genomes. This trend might be attributed to their prolonged adaptive evolution within parasitic environments (Ochman and Davalos, 2006).

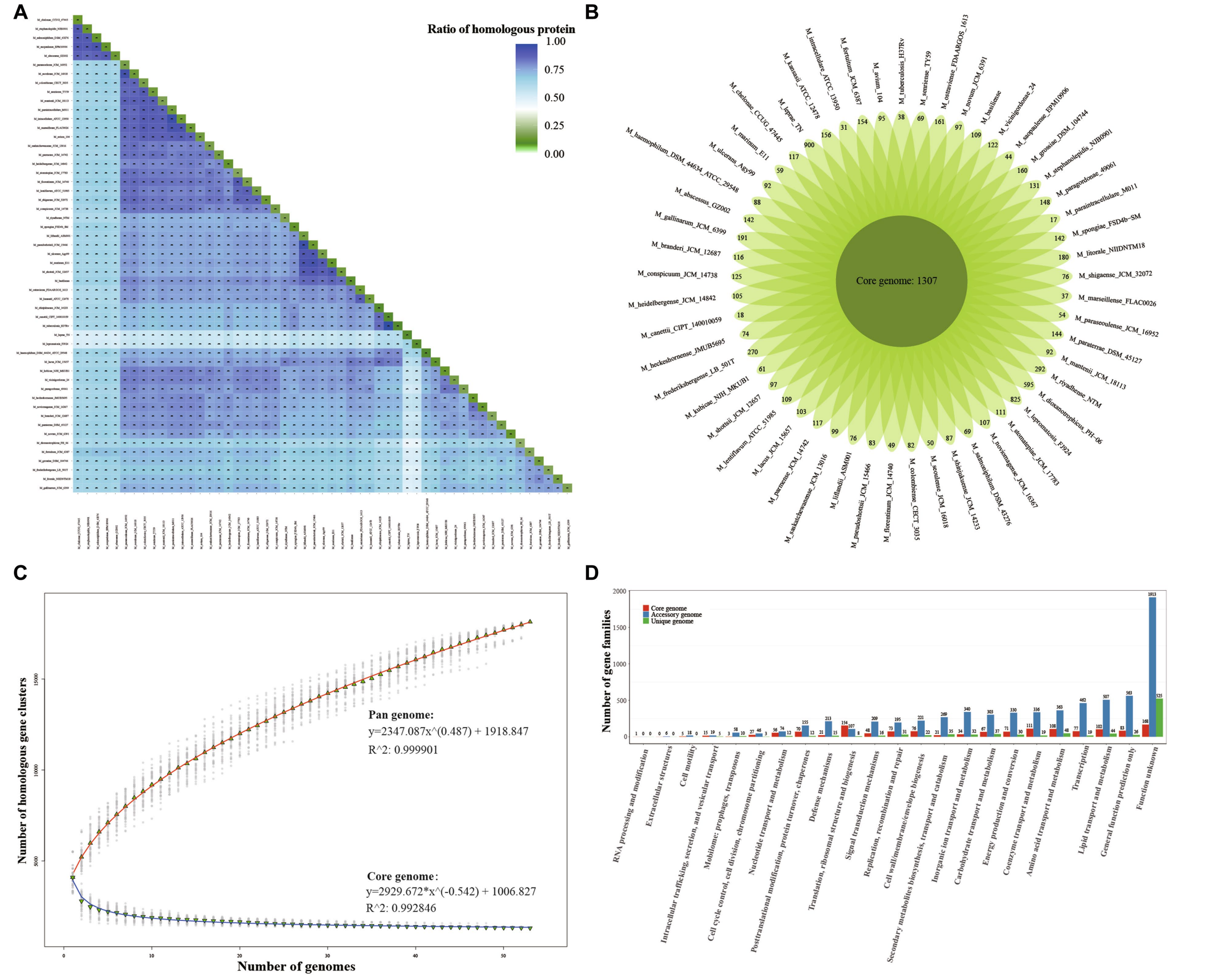


FIGURE 1
Core/Pan-genome structure and gene function distribution. **(A)** Percentage of homologous gene families for each pair of *Mycobacterium* species. The corresponding tiles display the percentage and number of homologous gene families, as well as the mean number of gene families, for each species pair. **(B)** Core genome size and unique gene sizes for each *Mycobacterium* species. **(C)** Cumulative size curve for core (blue) and pan-genome (red). The blue curve represents the cumulative sizes of the core genome, while the red curve depicts the cumulative sizes of the pan-genome. **(D)** COG categories of core, accessory, and unique genomes for *Mycobacterium*. The distribution of gene function categories within the core, accessory, and unique genomes of *Mycobacterium*.

From the outcomes of the analysis on homologous gene families, we identified a total of 18,139 gene families. Among these, 1,307 (7.21%) were shared across all 53 *Mycobacterium* species, 9,266 (51.08%) constituted the accessory genome, being shared by at least two but not all species, and the remaining 7,566 (41.71%) were characterized as singleton genes (Figure 1B). In essence, the core genes are accountable for the fundamental biological processes of a species and their key phenotypic traits. On the contrary, accessory and unique genes could potentially encode supplementary metabolic pathways that are not essential for survival but confer selective advantages (Medini et al., 2005; Tettelin et al., 2008). The prevalence of non-essential genomic content could potentially serve as the genetic foundation for *Mycobacterium* species to adapt to intricate environments and manifest diverse phenotypes. Interestingly, despite the small genomes of *M. leprae* and *M. lepromatosis*, they contained 900 and 825 singleton genes, respectively. This observation indicates distinct profiles in their genome composition. By means of curve fitting analysis, we derived the best-fitting function for core-genome size as $y = 2929.672^{*}\chi^{(-0.542)} + 1006.827$, and for pan-genome size as $y = 2347.087^{*}\chi^{(0.487)} + 1918.847$ (Figure 1C). Evidently, the core genome size reached a plateau after encompassing 30 species, whereas the pan-genome size exhibited rapid expansion as the number of species increased. This observation indicates that the *Mycobacterium* genus features an open pan-genome, highlighting a significant level of interspecific genetic diversity within *Mycobacterium*.

To explore whether the core, accessory, and unique genomes within the pan-genome exhibit distinct functional characteristics, we conducted an integrated functional annotation analysis using COG and EggNOG databases. As depicted in Figure 1D, the core-genome exhibited the two most enriched COG functional categories: "Translation, ribosomal structure, and biogenesis" (11.79%, 154/1307) and "Coenzyme transport and metabolism" (8.50%, 111/1307). In contrast, the proportions of these three categories within the accessory and unique genomes were 1.15, 3.62, and 0.11%, 0.25%, respectively. Regarding the accessory genome, the most enriched COG categories were "Lipid transport and metabolism" (5.47%) and "Transcription" (4.99%). This observation aligns with the anticipated notion that the core genome primarily encompasses housekeeping functions, as opposed to the accessory and unique genes. Furthermore, a notable prominence of "Mobilome: prophages, transposons," "Defense mechanisms," and "Signal transduction mechanisms" was evident within the accessory genome. These elements potentially contribute to their diverse niche adaptation and pathogenicity. However, it is noteworthy that 53.91% of accessory gene families and 94.81% of unique gene families lacked specific COG functional categorizations. This observation underscores the limited comprehension of gene function within *Mycobacterium* species, warranting further investigations in the future.

3.3 Expansion and contraction of gene families in *Mycobacterium* species

Remarkably, *Mycobacterium* species exhibit close genetic relationships while manifesting notable differences in phenotypes, including growth rate and pathogenicity (Pepperell, 2022). For a more comprehensive grasp of the evolutionary patterns among *Mycobacterium* species, we constructed a phylogenetic tree utilizing 709 single-copy core genes from the genomes of 53 *Mycobacterium* species. This was accomplished through the maximum likelihood method employing the JTT + F + R7 model (Figure 2A). The constructed tree substantiated the close relationship among all the fully pathogenic species, clustering them within the pathogenic mycobacteria (PT) node of the tree. This group encompassed *M. tuberculosis, M. leprae, M. marinum, M. ulcerans, M. canettii, M. shottsii* and others. *M. canettii* was identified as the species closest to *M. tuberculosis*. Notably, *M. ulcerans*, responsible for Buruli ulcer, formed a cluster alongside *M. shottsii, M. liflandii, M. pseudoshottsii*, and *M. marinum*—pathogens affecting fish and amphibians (Fremont-Rahl et al., 2011; Gauthier et al., 2022). These findings align with prior phylogenetic research grounded in average nucleotide identity (ANI) (Tortoli et al., 2017). A distinct demarcation exists between RGM and SGM species. All SGM species were positioned within the branches of the slow growth mycobacteria (SG) node on the tree. The *M. abscessus* complex constituted the most ancestral cluster of RGM, whereas *M. novum* emerged as the most ancestral species among SGM. The SGM group encompassed all TP species and a majority of the opportunity pathogenic species.

Gene family expansion and contraction constitute critical genetic factors influencing the habitat range and pathogenicity of bacteria (Baroncelli et al., 2016; Zhong et al., 2022). In this study, we reconstructed the genome-wide history of gene family contraction and expansion events across 53 *Mycobacterium* species. This reconstruction was conducted using the phylogenetic tree and encompassed all gene families within the pan-genome (Figure 2A). Our analysis revealed the participation of a total of 3,353 gene families (18.49%, 3,353/18,139) in contraction and expansion events. In the transition from RGM to SGM species, a majority of evolutionary events (77.78%) occurring at inner nodes involved gene family contractions. Specifically, there were 2,841 expansion events, encompassing 1,047 non-redundant gene families, alongside 9,947 contraction events, involving 2,580 non-redundant gene families, observed across the inner nodes. This outcome underscores the pivotal role of gene family contraction in the differentiation and formation of *Mycobacterium* species. Functional annotation using COG categories demonstrated the enrichment of "Transcription" and "Lipid transport and metabolism" within the gene families undergoing contraction and expansion events at inner nodes (Figure 2B). The regulation of gene expression stands as a pivotal mechanism for bacteria to adapt to intricate environments. It is widely recognized that the characteristic high lipid content, featuring diverse structures and biological activity, is a hallmark of *Mycobacterium* species (Lanéelle et al., 2021), contributing significantly to their survival and pathogenesis (Mallick et al., 2021). In contrast to inner nodes, the quantity of each COG category for contraction and expansion gene families within terminal branches was nearly uniform (Supplementary Figure S1). This variation in distribution patterns suggests that the gene families' contraction events play a significant role in the evolution and formation of *Mycobacterium* species.

3.4 Gene families associated with growth and pathogenicity traits in *Mycobacterium*

The evolutionary tree of *Mycobacterium* species distinctly separates RGM and SGM species, alongside TP, OP, and NP categories

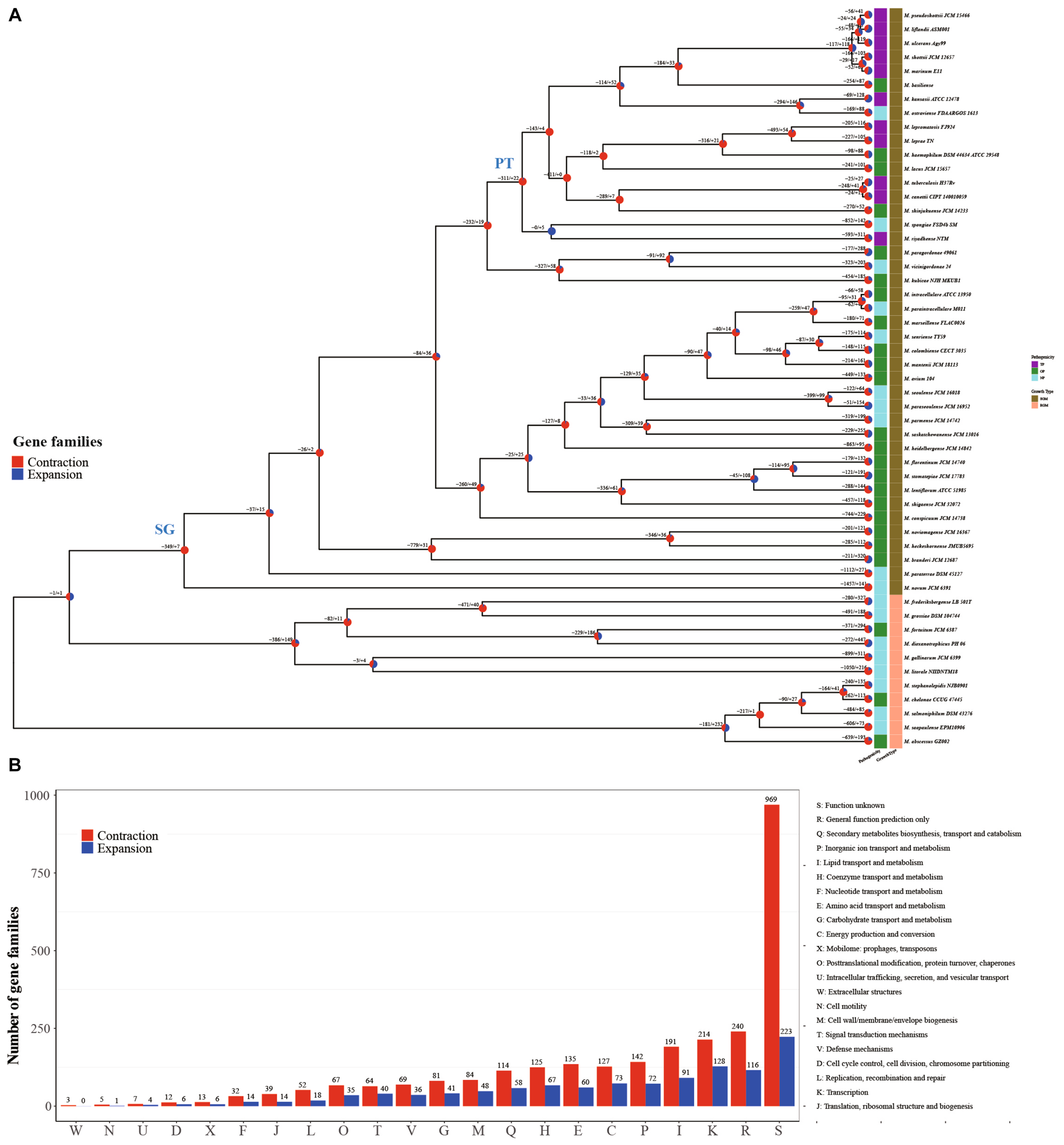


FIGURE 2
Structure and function distribution of gene family expansion and contraction in each *Mycobacterium* lineage. **(A)** Maximum-Likelihood tree constructed using concatenated amino acid sequences of 719 single-copy core genes. Each node on the tree is represented by a pie chart indicating the percentage of gene family contraction (red) and expansion (blue). The heatmap on the right of each species node depicts pathogenicity and growth type information. **(B)** Distribution of COG categories for contraction and expansion gene families of all inner nodes.

(Figure 2A). Crucially, the SG node and PT node served as pivotal points for the emergence of phenotypic alterations in growth rate and pathogenicity, respectively. From a logical standpoint, the contraction and expansion gene families associated with these nodes are intricately connected to the growth rate and pathogenicity characteristics of mycobacteria. In light of this, we meticulously conducted a comprehensive functional analysis of the contraction and expansion gene families linked to the SG and PT nodes. Virtually all of these events materialized as gene family contractions. To be specific, the counts of contraction and expansion gene families for the SG node were 349/7, whereas for the PT node, they stood at 311/22.

Regarding the SG node, the most prominently enriched COG categories among the contraction gene families included "E: Amino acid transport and metabolism" (8.82%, 31/349), "K: Transcription" (8.31%, 29/349), and "P: Inorganic ion transport and metabolism" (7.45%, 26/349). However, a notable proportion of the gene families (50.28%, 179/356) lacked definitive annotation results (Figure 3A, Supplementary Table S2). The protein–protein interaction (PPI)

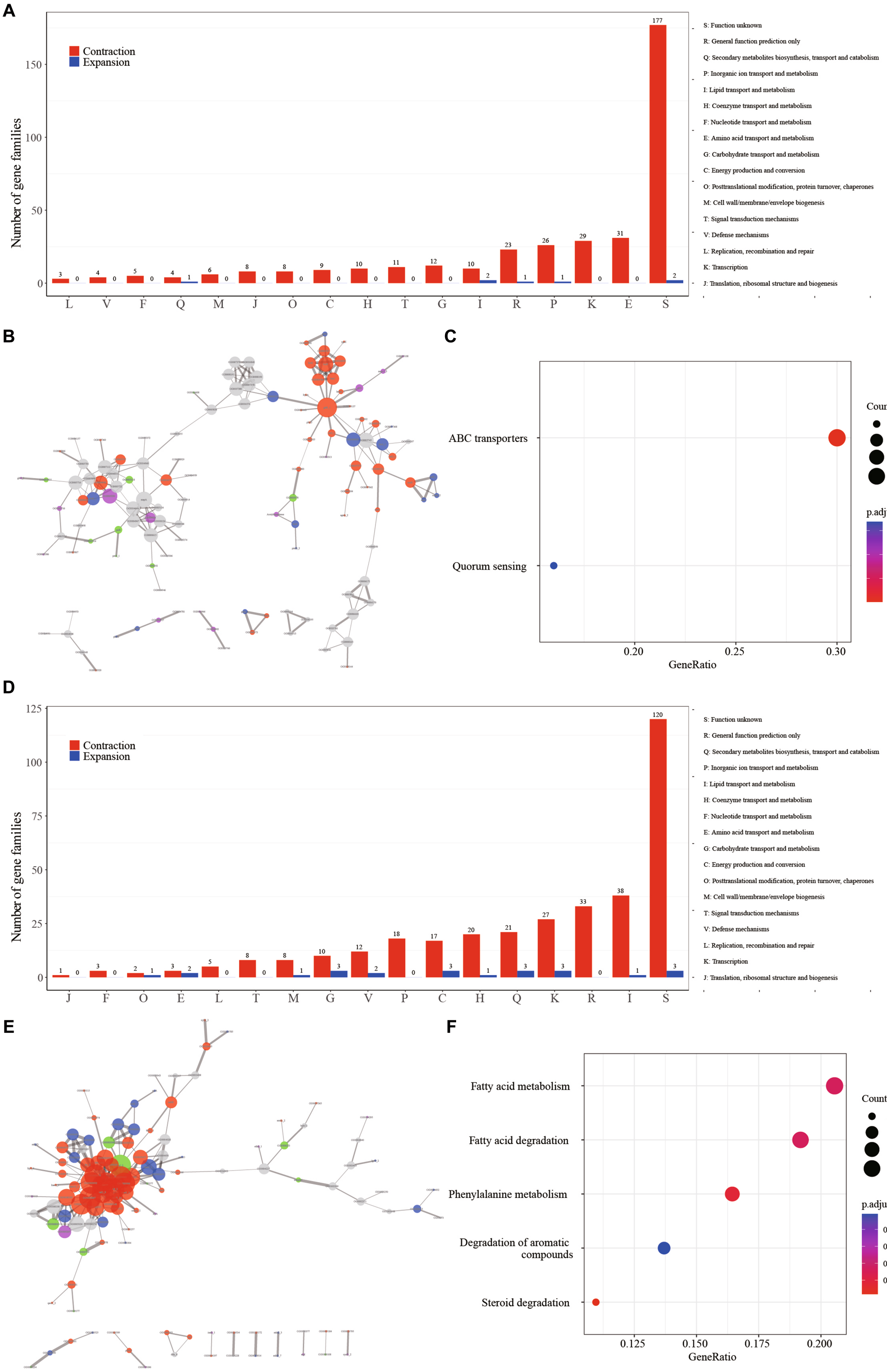


FIGURE 3

The major events in the context of growth speed type alteration and pathogenicity transition are characterized by the contraction of gene families. **(A)** Distribution of COG categories among contraction (depicted in red) and expansion (depicted in blue) gene families for GT node. **(B)** Protein–protein interaction network of contraction gene families for GT node. Network edges illustrate protein–protein interactions, with line thickness

(Continued)

FIGURE 3 (Continued)
indicating the confidence level of the relationships between nodes. Node size corresponds to the node degree. Nodes are color-coded to represent diverse protein function categories: red for basic substance transport and metabolism, purple for genetic information processing (including replication, transcription, and translation), blue for cellular processes (including cell wall/membrane/envelope biogenesis and cell motility), green for bacteria-environment interaction (including signal transduction, extracellular structures, and defense mechanisms), and grey for proteins lacking meaningful function annotation. **(C)** Enriched KEGG metabolism network of contraction gene families for GT node. **(D)** Distribution of COG categories among contraction (depicted in red) and expansion (depicted in blue) gene families for PT node. **(E)** Protein–protein interaction (PPI) network of contraction gene families for PT node. **(F)** Enriched KEGG metabolism network of contraction gene families for PT node.

network revealed the aggregation of these gene families into two distinct clusters (Figure 3B). A majority of these gene families were associated with fundamental bacterial substance transport and metabolic processes. Gene ontology (GO) enrichment analysis highlighted the substantial enrichment of gene families in ATP binding cassette (ABC) transporters and Quorum sensing (QS) (Figure 3C). Fifty gene families were implicated in ABC transporters, which play a pivotal role in bacterial physiology by facilitating the transport of diverse substrates impacting nutrition, pathogenesis, and antibiotic resistance (Orelle et al., 2019; Cassio Barreto De Oliveira and Balan, 2020). For instance, the ABC transporter encoded by the gene Rv1819c can transport unrelated hydrophilic compounds such as bleomycin and cobalamin, which are linked to the pathogenesis of tuberculosis (Rempel et al., 2020). The contraction of ABC transporter-related gene families in SGM could potentially diminish the efficiency of substance transport and metabolism, thereby contributing to their slow growth phenotypes. Facilitates biofilm formation, aiding bacterial survival in intricate environments (Sharma et al., 2014). Numerous *Mycobacterium* species, including *M. abscessus*, *M. avium*, *M. marinum,* and *M. fortuitum*, have demonstrated the ability to form biofilms across diverse conditions (Bardouniotis et al., 2003; Johansen et al., 2009). The contraction of gene families related to QS could signify an adaptive evolutionary response to a slow-growth state.

In contrast to the SG node, where gene contraction was predominantly associated with amino acid transport and metabolism, the PT node exhibited a distinct pattern. Here, gene families related to "I: Lipid transport and metabolism" (12.22%, 38/311), "R: Transcription (8.68%, 27/311)," and "Q: Secondary metabolites biosynthesis, transport and catabolism (6.75%, 21/311)" were prominently enriched, suggesting a strong link between these processes and pathogenicity (Figure 3D, Supplementary Table S3). The observation of these gene families clustering together in the protein–protein interaction network (Figure 3E) underscores their functional interconnectedness. This cluster primarily encompasses critical functions related to fundamental substance transport, metabolism, and cellular processes. Furthermore, the gene ontology (GO) enrichment analysis highlighted significant enrichment of five metabolic pathways in this gene cluster: "Fatty acid metabolism," "Fatty acid degradation," "Phenylalanine metabolism," "Degradation of aromatic compounds," and "Steroid degradation" (Figure 3F). Furthermore, the pivotal regulator FdmR in the pathogen *M. marinum* governs the abundance and chain length of virulence-associated lipids and mycolates, critically contributing to the maintenance of cell envelope impermeability. Notably, the role of fatty acid metabolism emerged as pivotal. In particular, the regulator *FdmR* in the pathogen *M. marinum* was highlighted for its control over the abundance and chain length of virulence-associated lipids and mycolates. This regulatory function is crucial for maintaining the impermeability of the cell envelope, a vital aspect of pathogenicity (Dong et al., 2021). Moreover, Yang et al. (2021) revealed that the hypoxia-induced mycobacterial protein, fatty-acid degradation A (FadA), plays a role in suppressing host immunity by modulating host fatty acid metabolism. This underscores the significance of fatty acid metabolism in mycobacteria's adaptation to the host's internal environment and in generating pathogenicity. Previous studies have demonstrated that members of the genus *Mycobacterium* are highly proficient in degrading aromatic compounds, which are prevalent environmental contaminants posing potential risks to human health (Kweon et al., 2015; Hennessee and Li, 2016). Interestingly, it's worth noting that mycobacteria, a genus known for its proficiency in degrading aromatic compounds, demonstrates a unique ecological adaptation (Kim et al., 2010). These compounds, common environmental contaminants with potential human health risks, are efficiently degraded by mycobacteria. However, it's intriguing that this associated gene family is notably absent in the majority of pathogenic bacteria. This raises questions about the ecological implications and potential genetic trade-offs associated with this adaptation in pathogenic versus non-pathogenic mycobacteria.

3.5 Variation in the presence and absence of virulence factors among *Mycobacterium* species

The genomes of pathogens harbor specific virulence factors that bestow the organism with the capacity to manipulate host immune defenses, enhance its disease-causing potential, and profoundly influence the course of infections (Bo et al., 2018; Shariq et al., 2023). Consequently, gaining a comprehensive grasp of these virulence factors holds paramount importance for attaining valuable insights into the mechanisms that underlie the infection process. The utilization of the pan-genome approach provides a distinctive avenue to pinpoint diverse pathogenic virulence genes dispersed across various *Mycobacterium* species and to probe the distribution or absence of these virulence genes within the *Mycobacterium* genus. Within the scope of this study, a total of 90 virulence genes were identified across the 53 *Mycobacterium* species, encompassing 26 core genes and 64 accessory genes (Figure 4). The functionalities of these virulence genes were associated with effector delivery systems and adherence (38 genes), nutritional and metabolic factors (19 genes), immune modulation (15 genes), regulation (9 genes), stress survival (4 genes), and other functions (2 genes) (Supplementary Table S4).

The core virulence genes primarily serve regulatory roles (*phoP, whiB3, VF0257, mprB, ideR, relA, sigE, hspR*) and participate in effector delivery systems (*PE5, esxG, esxH, eccE3, mycP3, eccD3, eccB3, eccC3, espG3, eccA3, espR*). Cluster analysis unveiled that pathogenic bacteria (Figure 4, cluster 5) exhibited the highest prevalence of virulence genes,

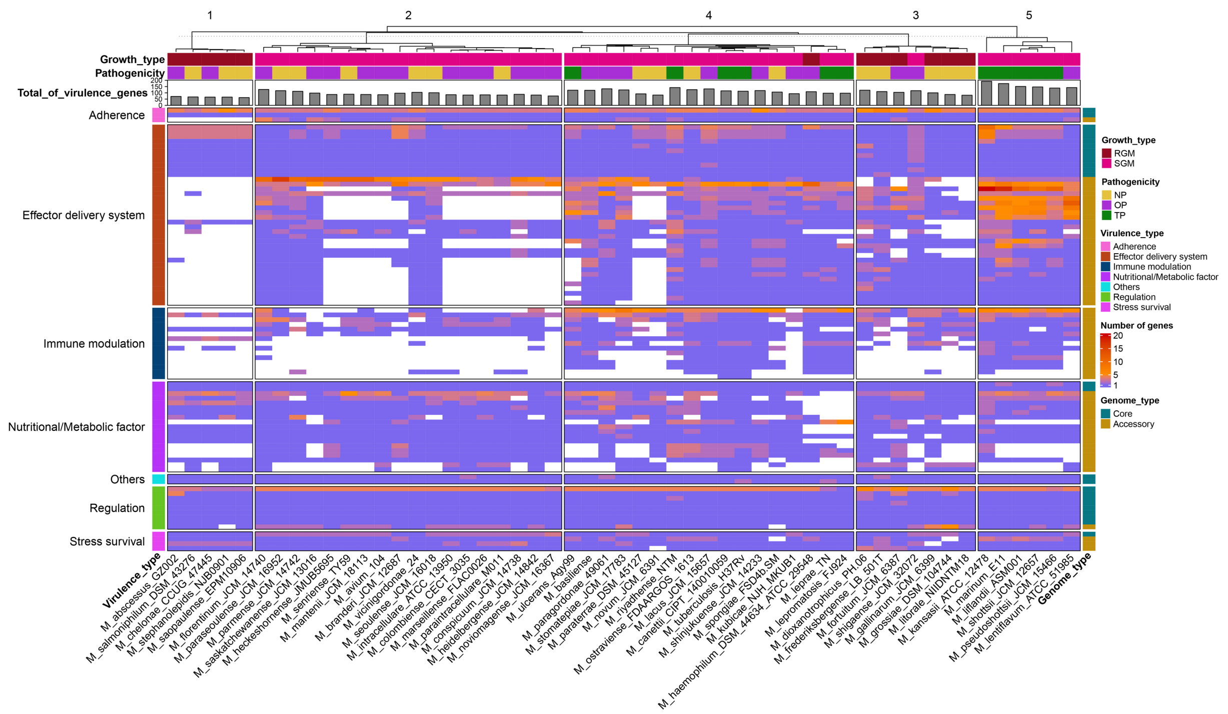


FIGURE 4
Presence-absence variation of virulence factors in each *Mycobacterium* species. The heatmap displays the number of virulence genes within corresponding gene families, represented by a color range, while white indicates the absence of the gene family. The virulence factors are color-coded based on VFDB classification on the left. Gene family types are indicated by color tiles on the right. Growth type and pathogenicity type annotations are provided as color tiles at the top of the heatmap. Additionally, the bar chart in the upper layer illustrates the total count of virulence genes for each *Mycobacterium* species.

whereas species belonging to the *M. abscessus-M. chelonae* complex (cluster 1) had the lowest presence. Notably, substantial variations in the distribution of accessory genes associated with immune modulation and effector delivery systems were observed between cluster 5 and the other four clusters. Phthiocerol dimycocerosate, a compound produced through the collaboration of *ddrB, drrC, fadD22, lppx, mmpL7, ppsA, ppsE, Rv2949c, Rv2951c, Rv2954c, Rv2958c, Rv2959c* and *tesA*, serves as a vital constituent of the mycobacterial cell wall, aiding in the evasion of host immune detection and counterattack (Rens et al., 2021). Mycobacteria showcase a multitude of paralogous type VII secretion systems known as Esx-1 to Esx-5. Notably, Esx-1 exerts a pivotal function in facilitating the *in vivo* expansion of pathogenic mycobacteria, and Esx-3 is essential for mycobactin-mediated iron acquisition as well as for *in vitro* growth (Siegrist et al., 2009, 2014). The aforementioned findings underscore that the distinct distribution patterns of these virulence genes stand as pivotal genetic underpinnings contributing to the diversity in pathogenicity observed among various *Mycobacterium* species.

M. tuberculosis and *M. leprae* exhibit a range of common attributes while presenting minimal distinctions, with notable variations observed primarily in growth rate and disease phenotypes (Hussain, 2007). While nearly all virulence factors are present in both *M. tuberculosis* and *M. canettii*, several Nutritional/Metabolic factors are notably absent in the *M. leprae* and *M. lepromatosis* (Supplementary Table S4). Notably, the genes *panC* and *panD*, which are pivotal in pantothenate biosynthesis, demonstrate conservation across all mycobacterial species (Figure 4). Pantothenic acid, also referred to as vitamin B5, plays a fundamental role in the biosynthesis of crucial molecules such as coenzyme A and acyl carrier protein (ACP) (Sambandamurthy et al., 2002). On the other hand, *M. leprae* and *M. lepromatosis* lack twelve genes responsible for encoding mycobactin, including *mbtA, mbtB, mbtD, mbtE, mbtF, mbtG, mbtH, mbtI, mbtJ, mbtK, mbtM* and *mbtN*. Reddy et al. (2013) have reported that the disruption of mycobactin biosynthesis leads to the attenuation of both the growth and virulence of *M. tuberculosis*. Thus, the absence of these genes in *M leprae* and *M. lepromatosis* could contribute to the significantly sluggish growth rate exhibited by these organisms.

There were 45 instances of expansion and contraction variation in virulence gene families across *Mycobacterium* species (Figure 5). Gene families associated with iron uptake factors and the Type VII secretion system exhibited reductions in rapid growth and non-pathogenic strains. However, these same gene families experienced substantial expansion in pathogenic or opportunistic pathogen strains. For instance, the gene families related to the ESAT-like secretion system 5 (ESX-5), including *EccA5, EccE5* and *EccD5*, expanded within cluster 4 (encompassing slow growth mycobacteria and pathogens), while undergoing contraction within cluster 1 (comprising rapid growth mycobacteria and non-pathogenic species). Deactivation of ESX-5 leads to a significant decrease in the secretion of PPE41, affecting the cell-to-cell migration of pathogenic mycobacteria (Abdallah et al., 2006). In *M. tuberculosis*, the EsxG-EsxH heterodimer has the ability to inhibit the host's endosomal sorting complex required for transport (ESCRT) machinery, suggesting that *M. tuberculosis* might impede this specific host response mechanism (Mittal et al., 2018). Notably, these

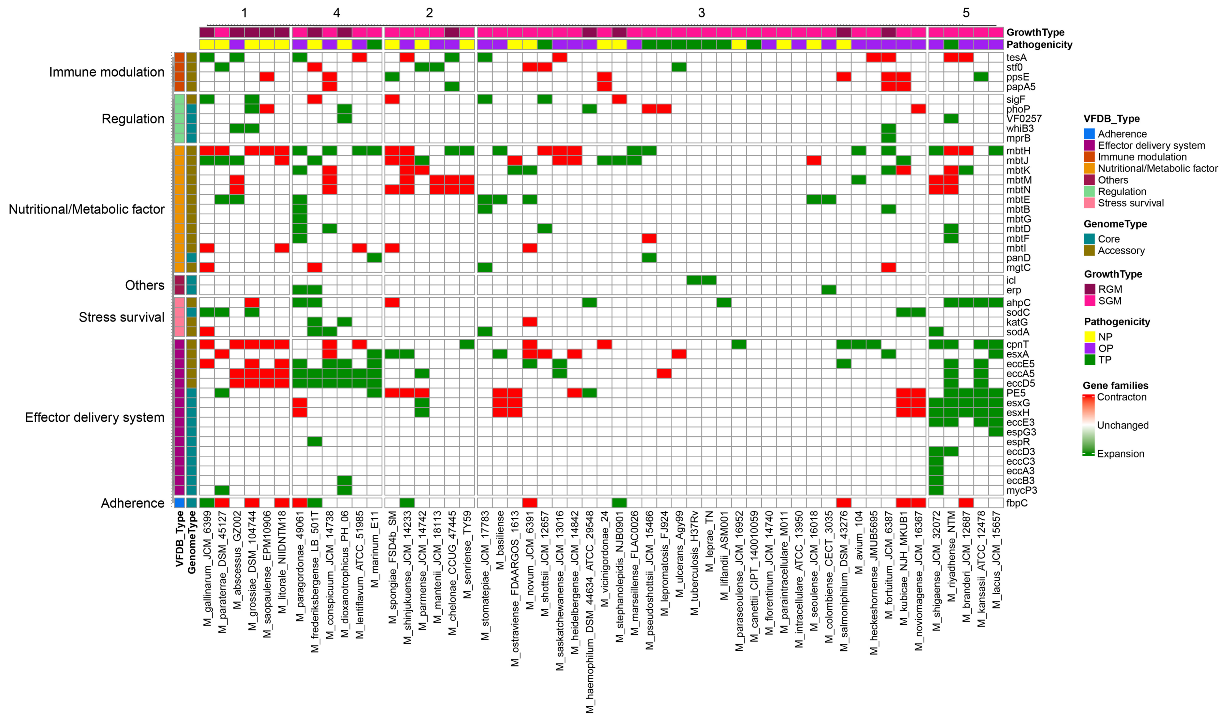


FIGURE 5
Variation in expansion and contraction of virulence gene families across *Mycobacterium* species. The heatmap illustrates the contraction (in red), expansion (in green), or unchanged status (in white) of virulence gene families for each mycobacterial species. The corresponding color tiles on the right denote the virulence and genome types of each gene family. Annotations for growth type and pathogenicity type are indicated at the top of the heatmap using color tiles.

gene families underwent expansion in several opportunistic pathogens, including *M. shinjukuense*, *M. riyadhense*, *M. branderi*, *M. kansasii* and *M. lacus*. These findings indicate that the expansion and contraction of virulence genes during the evolutionary process are tightly interconnected with the pathogenicity of the *Mycobacterium* strains.

3.6 Presence or absence variation of antibiotics resistance genes for *Mycobacterium* species

It is widely recognized that infections caused by *M. tuberculosis* and non-tuberculosis mycobacteria (NTM) are often challenging to treat effectively due to their multidrug-resistant properties (Zumla et al., 2014; Ratnatunga et al., 2020). Bacteria can acquire resistance to antibiotics through genetic mutations or the acquisition of resistance genes from other organisms. Therefore, understanding the distribution of antibiotic resistance genes within bacterial populations is crucial for effective disease management and treatment. To investigate the distribution of antibiotics resistance genes of 53 *Mycobacterium* species, the proteome of each mycobacteria were analyzed based CARD database.

As depicted in Supplementary Figure S2, our analysis revealed the presence of eleven antibiotic resistance genes associated with resistance to ten distinct drug classes, including rifamycin, isoniazid-like antibiotics, glycopeptide antibiotics, and others. These resistance genes play a pivotal role in the ability of *Mycobacterium* species to withstand the effects of various antibiotics. These genes are linked to four primary mechanisms of drug resistance, encompassing antibiotic efflux, inactivation, antibiotic target alteration, and protection. Particularly noteworthy is the widespread distribution of antibiotic resistance genes related to rifamycin and isoniazid (*efpA*, *rpoB*), macrolide and penam (*mtrA*), and phosphonic acid (*murA*) across the 53 mycobacterial species. Intriguingly, a consistent distribution of antimicrobial genes was observed among almost all TP species, with the exception of *M. leprae* and *M. lepromatosis*, as well as RGM species. This suggests that many drug-resistant genes within the *Mycobacterium* genus are inherent to the strains and are potentially reinforced by the intrinsic multidrug resistance conferred by the presence of a lipid-rich outer membrane (Nasiri et al., 2017). However, it is worth noting that *M. tuberculosis* and NTM exhibit significant heterogeneity in drug resistance (Gopalaswamy et al., 2020), highlighting the necessity of tailoring antibiotic treatment regimens based on strain phenotypes in clinical settings. Furthermore, the presence of specific resistance genes in non-pathogenic species raises intriguing questions about the potential sources and mechanisms of resistance gene dissemination among *Mycobacterium* species.

This study provides a comprehensive analysis of the expansion and contraction patterns in gene families of different *Mycobacterium* species and their association with growth rate and pathogenicity. The identified gene family expansion/contraction events related to pathogenicity, such as the type VII secretion systems and mycobactin biosynthesis, have significant implications for a deeper understanding of the pathogenic mechanisms of *M. tuberculosis*. Previous comparative genomics studies of *M. tuberculosis* have primarily focused on single nucleotide variations, horizontal gene transfers, and sequence-level changes (Gautam et al., 2017; Coll et al., 2018). In the long course of evolution, it is noteworthy to investigate whether different *M. tuberculosis* lineages/sublineages have experienced gene family expansion and contraction events related to their adaptability. This study serves as a reference for gene family analyses in *M. tuberculosis* lineages and other species. However, this study has certain limitations. First, *Mycobacterium* encompasses at least 190 species, and a more extensive dataset of *Mycobacterium* species genomes may yield more precise results. Furthermore, the findings of this study need further experimental validation in the future.

4 Conclusion

In conclusion, our study delved into the intricate genomic dynamics underlying growth rate variations and pathogenicity shifts within the *Mycobacterium* genus. Through an extensive analysis of 53 species, we have unveiled the genetic determinants responsible for these crucial phenotypic traits. The pan-genome analysis revealed a dynamic nature of the *Mycobacterium* genus, with core and accessory genomes playing distinct roles in maintaining essential functions and facilitating adaptability, respectively. Notably, the core genome appeared to stabilize after a certain number of species, while the pan-genome continued to expand, underscoring the genus' remarkable genetic diversity. Through the exploration of gene families associated with growth rate changes and pathogenicity shifts, we uncovered pivotal genetic determinants responsible for these phenotypic variations. Our findings underscore the significant role of gene family contractions, particularly those linked to nucleic acids, ions, and substance metabolism, in driving the intricate process of host adaptive evolution. Conversely, the expansion of virulence-associated gene families, notably encompassing the type VII secretion system and mycobactin biosynthesis, emerges as an equally pivotal determinant in shaping the pathogenicity landscape of mycobacteria. These dual mechanisms collectively highlight the dynamic interplay between genetic alterations and virulence factors that ultimately define the nuanced pathogenic potential of mycobacterial species. Moreover, our study delved into the intriguing landscape of antibiotic resistance genes, revealing their widespread presence across *Mycobacterium* species. The association of specific resistance genes with growth types and pathogenicity profiles highlighted their intrinsic nature in certain strains, potentially stemming from lipid-rich outer membranes.

Overall, this research provides valuable insights into the intricate interplay between genetic makeup, phenotypic traits, and pathogenicity in the *Mycobacterium* genus. These findings contribute to a deeper comprehension of the evolutionary dynamics that mold these bacterial populations. As we continue to decipher the complex interplay between genetics and phenotypic traits, we pave the way for potential advancements in diagnostics, therapies, and strategies for combating mycobacterial infections.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

XiaZ: Methodology, Writing – original draft, Investigation. QuL: Investigation, Writing – original draft, Data curation, Formal analysis. YL: Data curation, Investigation, Writing – review & editing. QiL: Data curation, Writing – review & editing. XinZ: Data curation, Writing – review & editing. XL: Writing – review & editing, Conceptualization. DC: Writing – review & editing, Conceptualization, Funding acquisition, Methodology, Supervision, Visualization, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Science and Technology Base and Talent Program in Guangxi (GuikeAD23026078). We also thanks for the support of Construction of Key Clinical Specialties in Guangxi (Guiweiyifa [2022]-NO.21) and the Key Laboratory of Molecular Pathology of Guangxi.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1292897/full#supplementary-material

References

Abdallah, A. M., Verboom, T., Hannes, F., Safi, M., Strong, M., Eisenberg, D., et al. (2006). A specific secretion system mediates Ppe41 transport in pathogenic mycobacteria. *Mol. Microbiol.* 62, 667–679. doi: 10.1111/j.1365-2958.2006.05409.x

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2019). Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 48, D517–D525. doi: 10.1093/nar/gkz935

Bachmann, N. L., Salamzade, R., Manson, A. L., Whittington, R., Sintchenko, V., Earl, A. M., et al. (2019). Key transitions in the evolution of rapid and slow growing identified by comparative genomics. *Front. Microbiol.* 10:3019. doi: 10.3389/fmicb.2019.03019

Bardouniotis, E., Ceri, H., and Olson, M. E. (2003). Biofilm formation and biocide susceptibility testing of mycobacterium fortuitum and *Mycobacterium marinum*. *Curr. Microbiol.* 46, 28–32. doi: 10.1007/s00284-002-3796-4

Baroncelli, R., Amby, D. B., Zapparata, A., Sarrocco, S., Vannacci, G., Le Floch, G., et al. (2016). Gene family expansions and contractions are associated with host range in plant pathogens of the genus Colletotrichum. *BMC Genomics* 17:555. doi: 10.1186/s12864-016-2917-6

Bartlett, A., Padfield, D., Lear, L., Bendall, R., and Vos, M. (2022). A comprehensive list of bacterial pathogens infecting humans. *Microbiology* 168:1269. doi: 10.1099/mic.0.001269

Bo, L., Dandan, Z., Qi, J., Lihong, C., and Jian, Y. (2018). Vfdb 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 47, D687–D692. doi: 10.1093/nar/gky1080

Cassio Barreto De Oliveira, M., and Balan, A. (2020). The Atp-binding cassette (Abc) transport Systems in *Mycobacterium tuberculosis*: structure, function, and possible targets for therapeutics. *Biology* 9:443. doi: 10.3390/biology9120443

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544. doi: 10.1038/31159

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 50, 307–316. doi: 10.1038/s41588-017-0029-0

Dong, W., Nie, X., Zhu, H., Liu, Q., Shi, K., You, L., et al. (2021). Mycobacterial fatty acid catabolism is repressed by FdmR to sustain lipogenesis and virulence. *Proc. Natl. Acad. Sci.* 118:e2019305118. doi: 10.1073/pnas.2019305118

Dulberger, C. L., Rubin, E. J., and Boutte, C. C. (2020). The mycobacterial cell envelope — a moving target. *Nat. Rev. Microbiol.* 18, 47–59. doi: 10.1038/s41579-019-0273-7

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238. doi: 10.1186/s13059-019-1832-y

Fremont-Rahl, J. J., Ek, C., Williamson, H. R., Small, P. L., Fox, J. G., and Muthupalani, S. (2011). Mycobacterium liflandii outbreak in a research colony of Xenopus (Silurana) tropicalis frogs. *Vet. Pathol.* 48, 856–867. doi: 10.1177/0300985810388520

Gautam, S. S., Mac Aogáin, M., Bower, J. E., Basu, I., and O'toole, R. F. (2017). Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of *Mycobacterium tuberculosis*. *Infect. Dis.* 49, 680–688. doi: 10.1080/23744235.2017.1330553

Gauthier, D. T., Doss, J. H., Lagatta, M., Gupta, T., Karls, R. K., and Quinn, F. D. (2022). Genomic degeneration and reduction in the fish pathogen *Mycobacterium shottsii*. *Microbiol. Spectr.* 10:e0115821. doi: 10.1128/spectrum.01158-21

Gauthier, D. T., Haines, A. N., and Vogelbein, W. K. (2021). Elevated temperature inhibits *Mycobacterium shottsii* infection and *Mycobacterium pseudoshottsii* disease in striped bass *Morone saxatilis*. *Dis. Aquat. Org.* 144, 159–174. doi: 10.3354/dao03584

Gopalaswamy, R., Shanmugam, S., Mondal, R., and Subbian, S. (2020). Of tuberculosis and non-tuberculous mycobacterial infections – a comparative analysis of epidemiology, diagnosis and treatment. *J. Biomed. Sci.* 27:74. doi: 10.1186/s12929-020-00667-6

Gu, Z. (2022). Complex heatmap visualization. *iMeta* 1:e43

Hennessee, C. T., and Li, Q. X. (2016). Effects of polycyclic aromatic hydrocarbon mixtures on degradation, gene expression, and metabolite production in four mycobacterium species. *Appl. Environ. Microbiol.* 82, 3357–3369. doi: 10.1128/AEM.00100-16

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). Eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Hussain, T. (2007). Leprosy and tuberculosis: an insight-review. *Crit. Rev. Microbiol.* 33, 15–66. doi: 10.1080/10408410601172271

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Johansen, T. B., Agdestein, A., Olsen, I., Nilsen, S. F., Holstad, G., and Djønne, B. (2009). Biofilm formation by *Mycobacterium avium* isolates originating from humans, swine and birds. *BMC Microbiol.* 9:159. doi: 10.1186/1471-2180-9-159

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kim, S. J., Kweon, O., and Cerniglia, C. E. (2010). “Degradation of polycyclic aromatic hydrocarbons by mycobacterium strains” in *Handbook of hydrocarbon and lipid microbiology*. ed. K. N. Timmis (Berlin, Heidelberg: Springer Berlin Heidelberg)

Kweon, O., Kim, S.-J., Blom, J., Kim, S.-K., Kim, B.-S., Baek, D.-H., et al. (2015). Comparative functional pan-genome analyses to build connections between genomic dynamics and phenotypic evolution in polycyclic aromatic hydrocarbon metabolism in the genus mycobacterium. *BMC Evol. Biol.* 15:21. doi: 10.1186/s12862-015-0302-8

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). Rnammer: consistent and rapid annotation of ribosomal Rna genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Lanéelle, M. A., Spina, L., Nigou, J., Lemassu, A., and Daffé, M. (2021). Lipid and Lipoarabinomannan isolation and characterization. *Methods Mol. Biol.* 2314, 109–150. doi: 10.1007/978-1-0716-1460-0_4

Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). Vfdb 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50, D912–d917. doi: 10.1093/nar/gkab1107

Lu, Q., Zhu, X., Long, Q., Yi, X., Yang, A., Long, X., et al. (2022). Comparative genomics reveal the utilization ability of variable carbohydrates as key genetic features of listeria pathogens in their pathogenic lifestyles. *Pathogens* 11:1430. doi: 10.3390/pathogens11121430

Mallick, I., Santucci, P., Poncin, I., Point, V., Kremer, L., Cavalier, J. F., et al. (2021). Intrabacterial lipid inclusions in mycobacteria: unexpected key players in survival and pathogenesis? *FEMS Microbiol. Rev.* 45:fuab029. doi: 10.1093/femsre/fuab029

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594. doi: 10.1016/j.gde.2005.09.006

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). Cafe 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36, 5516–5518. doi: 10.1093/bioinformatics/btaa1022

Mittal, E., Skowyra, M. L., Uwase, G., Tinaztepe, E., Mehra, A., Köster, S., et al. (2018). *Mycobacterium tuberculosis* type vii secretion system effectors differentially impact the Escrt endomembrane damage response. *MBio* 9:e01765-18. doi: 10.1128/mbio.01765-18

Nasiri, M. J., Haeili, M., Ghazi, M., Goudarzi, H., Pormohammad, A., Imani Fooladi, A. A., et al. (2017). New insights in to the intrinsic and acquired drug resistance mechanisms in mycobacteria. *Front. Microbiol.* 8:681. doi: 10.3389/fmicb.2017.00681

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Ochman, H., and Davalos, L. M. (2006). The nature and dynamics of bacterial genomes. *Science* 311, 1730–1733. doi: 10.1126/science.1119966

Orelle, C., Mathieu, K., and Jault, J.-M. (2019). Multidrug Abc transporters in bacteria. *Res. Microbiol.* 170, 381–391. doi: 10.1016/j.resmic.2019.06.001

Pennington, K. M., Vu, A., Challener, D., Rivera, C. G., Shweta, F. N. U., Zeuli, J. D., et al. (2021). Approach to the diagnosis and treatment of non-tuberculous mycobacterial disease. *J. Clin. Tuberc. Other Mycobact. Dis.* 24:100244. doi: 10.1016/j.jctube.2021.100244

Pepperell, C. S. (2022). Evolution of tuberculosis pathogenesis. *Annu. Rev. Microbiol.* 76, 661–680. doi: 10.1146/annurev-micro-121321-093031

Philley, J. V., and Griffith, D. E. (2015). Treatment of slowly growing mycobacteria. *Clin. Chest Med.* 36, 79–90. doi: 10.1016/j.ccm.2014.10.005

Ratnatunga, C. N., Lutzky, V. P., Kupz, A., Doolan, D. L., Reid, D. W., Field, M., et al. (2020). The rise of non-tuberculosis mycobacterial lung disease. *Front. Immunol.* 11. doi: 10.3389/fimmu.2020.00303

Reddy, P. V., Puri, R. V., Chauhan, P., Kar, R., Rohilla, A., Khera, A., et al. (2013). Disruption of mycobactin biosynthesis leads to attenuation of *Mycobacterium tuberculosis* for growth and virulence. *J. Infect. Dis.* 208, 1255–1265. doi: 10.1093/infdis/jit250

Rempel, S., Gati, C., Nijland, M., Thangaratnarajah, C., Karyolaimos, A., De Gier, J. W., et al. (2020). A mycobacterial Abc transporter mediates the uptake of hydrophilic compounds. *Nature* 580, 409–412. doi: 10.1038/s41586-020-2072-8

Rens, C., Chao, J. D., Sexton, D. L., Tocheva, E. I., and Av-Gay, Y. (2021). Roles for phthiocerol dimycocerosate lipids in *Mycobacterium tuberculosis* pathogenesis. *Microbiology* 167:1042. doi: 10.1099/mic.0.001042

Sambandamurthy, V. K., Wang, X., Chen, B., Russell, R. G., Derrick, S., Collins, F. M., et al. (2002). A pantothenate auxotroph of *Mycobacterium tuberculosis* is highly attenuated and protects mice against tuberculosis. *Nat. Med.* 8, 1171–1174. doi: 10.1038/nm765

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shariq, M., Quadir, N., Alam, A., Zarin, S., Sheikh, J. A., Sharma, N., et al. (2023). The exploitation of host autophagy and ubiquitin machinery by *Mycobacterium tuberculosis* in shaping immune responses and host defense during infection. *Autophagy* 19, 3–23. doi: 10.1080/15548627.2021.2021495

Sharma, I. M., Petchiappan, A., and Chatterji, D. (2014). Quorum sensing and biofilm formation in mycobacteria: role of c-di-Gmp and methods to study this second messenger. *IUBMB Life* 66, 823–834. doi: 10.1002/iub.1339

Siegrist, M. S., Steigedal, M., Ahmad, R., Mehra, A., Dragset, M. S., Schuster, B. M., et al. (2014). Mycobacterial Esx-3 requires multiple components for iron acquisition. *mBio,* 5. doi: 10.1128/mbio.01073-14,

Siegrist, M. S., Unnikrishnan, M., Mcconnell, M. J., Borowsky, M., Cheng, T. Y., Siddiqi, N., et al. (2009). Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 18792–18797. doi: 10.1073/pnas.0900589106

Steenwyk, J. L., Buida, T. J. 3rd, Li, Y., Shen, X. X., and Rokas, A. (2020). Clipkit: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* 18:e3001007. doi: 10.1371/journal.pbio.3001007

Supply, P., Marceau, M., Mangenot, S., Roche, D., Rouanet, C., Khanna, V., et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat. Genet.* 45, 172–179. doi: 10.1038/ng.2517

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–d612. doi: 10.1093/nar/gkaa1074

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006

Tortoli, E., Fedrizzi, T., Meehan, C. J., Trovato, A., Grottola, A., Giacobazzi, E., et al. (2017). The new phylogeny of the genus mycobacterium: the old and the news. *Infect. Genet. Evol.* 56, 19–25. doi: 10.1016/j.meegid.2017.10.013

Tufariello, J. M., Chapman, J. R., Kerantzas, C. A., Wong, K. W., Vilchèze, C., Jones, C. M., et al. (2016). Separable roles for *Mycobacterium tuberculosis* Esx-3 effectors in iron acquisition and virulence. *Proc. Natl. Acad. Sci. U. S. A.* 113, E348–E357. doi: 10.1073/pnas.1523321113

Tullius, M. V., Nava, S., and Horwitz, M. A. (2019). Ppe37 is essential for *Mycobacterium tuberculosis* Heme-iron acquisition (Hia), and a defective Ppe37 in *Mycobacterium bovis* Bcg prevents Hia. *Infect. Immun.* 87:e00540-18. doi: 10.1128/IAI.00540-18

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, H., Wang, F., Guo, X., Liu, F., Liu, Z., Wu, X., et al. (2021). Interception of host fatty acid metabolism by mycobacteria under hypoxia to suppress anti-Tb immunity. *Cell Discov.* 7:90. doi: 10.1038/s41421-021-00301-1

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. (2016). Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zhong, C., Qu, B., Hu, G., and Ning, K. (2022). Pan-genome analysis of campylobacter: insights on the genomic diversity and virulence profile. *Microbiol. Spectr.* 10, e01029–e01022. doi: 10.1128/spectrum.01029-22

Zumla, A. I., Gillespie, S. H., Hoelscher, M., Philips, P. P. J., Cole, S. T., Abubakar, I., et al. (2014). New antituberculosis drugs, regimens, and adjunct therapies: needs, advances, and future prospects. *Lancet Infect. Dis.* 14, 327–340. doi: 10.1016/S1473-3099(13)70328-1

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 05 January 2024
DOI 10.3389/fmicb.2023.1277468

Check for updates

OPEN ACCESS

EDITED BY
Augusto Simoes-Barbosa,
The University of Auckland, New Zealand

REVIEWED BY
Siva Uppalapati,
University of Colorado Anschutz Medical Campus, United States
Mikeljon P. Nikolich,
Walter Reed Army Institute of Research, United States

*CORRESPONDENCE
Isabel Lopes de Carvalho
✉ isabel.carvalho@insa.min-saude.pt

†These authors share first authorship

RECEIVED 14 August 2023
ACCEPTED 23 November 2023
PUBLISHED 05 January 2024

CITATION
Isidro J, Escudero R, Luque-Larena JJ, Pinto M, Borges V, González-Martín-Niño R, Duarte S, Vieira L, Mougeot F, Vidal D, Herrera -Rodríguez D, Rodríguez-Pastor R, Herrero-Cófreces S, Jubete-Tazo F, Gomes JP and Lopes de Carvalho I (2024) Strengthening the genomic surveillance of *Francisella tularensis* by using culture-free whole-genome sequencing from biological samples.
Front. Microbiol. 14:1277468.
doi: 10.3389/fmicb.2023.1277468

Strengthening the genomic surveillance of *Francisella tularensis* by using culture-free whole-genome sequencing from biological samples

Joana Isidro[1†], Raquel Escudero[2†], Juan José Luque-Larena[3], Miguel Pinto[1], Vítor Borges[1], Rosa González-Martín-Niño[2], Sílvia Duarte[4], Luís Vieira[4], François Mougeot[5], Dolors Vidal[6], Daniel Herrera-Rodríguez[5,6], Ruth Rodríguez-Pastor[7,8], Silvia Herrero-Cófreces[3], Fernando Jubete-Tazo[3], João Paulo Gomes[1,9] and Isabel Lopes de Carvalho[10]*

[1]Genomics and Bioinformatics Unit, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal, [2]Reference and Research Laboratory on Special Pathogens, National Centre for Microbiology (CNM), Carlos II Health Institute (ISCIII), Madrid, Spain, [3]Departamento de Ciencias Agroforestales, Instituto Universitario de Investigación en Gestión Forestal Sostenible (iuFOR), E.T.S. Ingenierías Agrarias, Universidad de Valladolid, Palencia, Spain, [4]Technology and Innovation Unit, Department of Human Genetics, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal, [5]Instituto de Investigación en Recursos Cinegéticos (IREC-CSIC, UCLM, JCCM), Ciudad Real, Spain, [6]Área de Microbiología, Facultad de Medicina, Universidad de Catilla-La Mancha (UCLM), Ciudad Real, Spain, [7]Department of Parasitology, Veterinary Faculty, University of Zaragoza, Zaragoza, Spain, Zaragoza, Spain, [8]Departamento de Parasitología, Facultad de Veterinaria, Universidad de Zaragoza, Instituto Agroalimentario de Aragón-IA2 (Universidad de Zaragoza-CITA), Zaragoza, Spain, [9]Veterinary and Animal Research Center (CECAV), Faculty of Veterinary Medicine, Lusófona University, Lisbon, Portugal, [10]Emergency and Biopreparedness Unit, National Institute of Health Doutor Ricardo Jorge, Lisbon, Portugal

Introduction: *Francisella tularensis* is a highly infectious bacterium that causes the zoonotic disease tularemia. The development of genotyping methods, especially those based on whole-genome sequencing (WGS), has recently increased the knowledge on the epidemiology of this disease. However, due to the difficulties associated with the growth and isolation of this fastidious pathogen in culture, the availability of strains and subsequently WGS data is still limited.

Methods: To surpass these constraints, we aimed to implement a culture-free approach to capture and sequence *F. tularensis* genomes directly from complex samples. Biological samples obtained from 50 common voles and 13 Iberian hares collected in Spain were confirmed as positive for *F. tularensis subsp. holarctica* and subjected to a WGS target capture and enrichment protocol, using RNA oligonucleotide baits designed to cover *F. tularensis* genomic diversity.

Results: We obtained full genome sequences of *F. tularensis* from 13 animals (20.6%), two of which had mixed infections with distinct genotypes, and achieved a higher success rate when compared with culture-dependent WGS (only successful for two animals). The new genomes belonged to different clades commonly identified in Europe (B.49, B.51 and B.262) and subclades. Despite being phylogenetically closely related to other genomes from Spain, the detected clusters were often found in other countries. A comprehensive phylogenetic analysis, integrating 599 *F. tularensis* subsp.

holarctica genomes, showed that most (sub)clades are found in both humans and animals and that closely related strains are found in different, and often geographically distant, countries.

Discussion: Overall, we show that the implemented culture-free WGS methodology yields timely, complete and high-quality genomic data of *F. tularensis*, being a highly valuable approach to promote and potentiate the genomic surveillance of *F. tularensis* and ultimately increase the knowledge on the genomics, ecology and epidemiology of this highly infectious pathogen.

KEYWORDS

***Francisella tularensis*, WGS, RNA oligonucleotide baits, *Microtus arvalis*, SureSelect**

1 Introduction

Francisella tularensis is a facultative intracellular bacterium causing the zoonotic disease tularemia. It is considered a potential bioterrorist agent due to its very low infective dose and considerable stability in aerosols, being classified as a Category A biological agent by Centers for Disease Control and Prevention (CDC) (n.d.). *F. tularensis* may be transmitted to humans by a number of different routes, including handling infected animals, ingestion of contaminated food or water, inhalation of infective aerosols or arthropod bites (ticks and insects) (Petersen et al., 2009). In nature, *F. tularensis* has been detected in many wild species including lagomorphs, rodents, insectivores, carnivores, ungulates, marsupials, birds, amphibians, fish, and invertebrates. In Northwest (NW) Spain, the region of Castilla-y-León is an endemic hotspot of the disease in southern Europe, accumulating more than 1,000 human cases since 1997 and its epidemiology is mainly associated with population outbreaks of common voles, *Microtus arvalis*, in intensive farming landscapes (Luque-Larena et al., 2017; Herrero-Cófreces et al., 2021).

There are two clinically relevant subspecies of *F. tularensis*: *F. tularensis* subsp. *tularensis* and *F. tularensis* subsp. *holarctica*. Only the less pathogenic subspecies *holarctica* has been detected in Europe (Dennis et al., 2001; Carvalho et al., 2014). The diagnosis of tularemia is complex due both to the nonspecific nature of the initial symptoms and to the fact that *F. tularensis* is difficult to culture (Doern, 2000), showing slow growth rates, with individual colonies on nonselective optimized agar plates usually requiring two to four days of incubation to be visible (Aloni-Grinstein et al., 2017).

Environmental studies on the distribution of *Francisella* spp. are hampered by the frequency of other species of *Francisella* such as *Francisella*-like endosymbionts and other so far not proven pathogenic species that can produce a misleading positive result (Escudero et al., 2008). Efforts have been made to improve methods for the discrimination on *Francisella* species.

Considering the clonal nature of this species, only molecular methods with high discriminatory power, such as whole genome sequencing (WGS), allow to distinguish between closely related subpopulations at strain level (Johansson and Petersen, 2010; Dwibedi et al., 2016). The genotyping of *F. tularensis* strains is mostly based on the identification of canonical SNPs defined on a whole-genome level (Kevin et al., 2020). However, despite rapidly accumulating knowledge, the phylogeography of the pathogen is still poorly understood due to the low availability of isolates for WGS and consequent lack of *F. tularensis* genomic data in many tularemia-endemic countries (Shevtsov et al., 2021).

In this study, we implemented a culture-free approach for capturing and sequencing *F. tularensis* genomes directly from complex biological samples, passing the constraints associated with the isolation and culture of this fastidious pathogen. Ultimately, we aimed to increase the knowledge on *F. tularensis* genomic epidemiology not only in Spain but also on a global scale by an integrative analysis with genomes from multiple countries.

2 Methods

2.1 Sample collection and selection

In NW Spain, the common vole is a key host and spillover agent of tularemia (Herrero-Cófreces et al., 2021). Vole samples were obtained from long-term ecological studies in which animals are captured in the field (Rodríguez-Pastor et al., 2017). Voles were live-trapped using Sherman© traps (8 cm × 9 cm × 23 cm, LFAHD Sherman©) baited with carrots and apples, which were set open in the morning and retrieved 24 h later. Captures were performed every four months (March, July and November) between July 2009 and July 2015. Captured voles were taken alive to the laboratory and euthanized using a CO_2 cabinet, following a protocol approved by the ethics committee from the University of Valladolid (authorization code: 4801646). Iberian hare samples were obtained from hunters and collected in the field during two hunting seasons (from October 2016 to February 2018). Subsequently, necropsies were performed in a microbiological safety cabinet, collecting different samples in aseptic conditions (mainly liver and spleen) for the detection of *F. tularensis*. All animals were captured from the provinces of Palencia (42°01′N, 4°42′W), Valladolid (41°34′N, 5°14′W) and Zamora (41°50′N, 5°36′W).

In the present study, the selected biological samples (liver, spleen and lung) were obtained from 50 *M. arvalis* that had tested positive for *F. tularensis* by real time-PCR, mostly with low Ct results. Moreover, DNAs extracted from homogenates of spleen and liver from 13 Iberian hares were tested as well, with signs suggestive of tularemia (i.e., hepatomegaly, splenomegaly or foci of necrosis in the selected tissues).

2.2 Culture of the samples

To culture the etiological agent, samples from 50 voles, consisting of lung, liver and spleen of each animal, were inoculated separately in culture media. Each tissue was punctured 10 to 20 times with a sterile wooden stick. Tissue adhering to the stick was transferred to chocolate agar PolyViteX (Biomerieux) and then a 1-μL sterile loop was used to streak the plate for colony isolation. All plates were incubated at 37°C for 5 more days and checked daily for characteristic *F. tularensis* growth (Petersen et al., 2004).

2.3 DNA extraction and real-time PCR

For cultures, DNA was extracted and purified using QIAGEN® Genomic-tip 100/G columns and QIAGEN® Genomic DNA Buffer sets from suspensions calibrated to 3 McFarland units according to the manufacturer's recommendations. For the biological samples, DNA from a homogenized mixture of liver and spleen (≈25 mg) was extracted by standard procedures using the QIAamp DNA Mini Kit (QIAGEN, Valencia, CA, United States).

Samples were tested using a real-time multitarget TaqMan PCR, using *tul4* and *ISFtu2* assays, and genome equivalents (GEs) concentration was inferred as previously described (Versage et al., 2003). Positive samples were further tested using real-time TaqMan PCR assays which differentiate between *F. tularensis* subsp. *tularensis* (type A) and *F. tularensis* subsp. *holarctica* (type B) (Kugeler et al., 2006). Moreover, a phylogenetically informative region of lipoprotein A (*lpnA*) gene (231 bp) was amplified by conventional PCR and hybridized with specific probes by reverse-line blotting (RLB) as previously described (Escudero et al., 2008).

For real-time PCR using tul4, ISFtu2, type A and type B assays, a type B positive control was used, as type A strains are restricted to North America.

In the four positive cultures of *F. tularensis*, the isolates were subjected to whole-genome sequencing on an Illumina MiSeq platform after DNA extraction and library preparation using the Nextera XT Library Preparation kit (Illumina, San Diego, CA, United States), according to manufacturer's instruction.

2.4 SureSelect$^{XT HS}$ custom library design for targeted sequencing of *Francisella tularensis*

RNA oligonucleotide baits (120 bp in size, a total of 54,756 baits) contiguously spanning (i.e., no tiling) the entire genome of *Francisella tularensis* were designed based on representative publicly available genomes, and plasmids, of the four *F. tularensis* subspecies (accession numbers NC_010677.1, NZ_CP009694.1, NZ_CP009683.1, NZ_CP010104.1, NZ_CP009352.1, NZ_CP010447.1, NC_006570.2, and NZ_CP009633.1). To ensure specificity, we excluded baits with considerable homology (determined by BLASTn) with the human (Human G + T) or mouse (Mouse G + T) genomes and transcriptomes or with available nucleotide sequences from the genus *Microtus* (taxonomy ID 10053), *Oryctolagus* (taxonomy ID 9984) or *Lepus* (taxonomy ID 9980). Finally, this custom library was uploaded to Agilent's SureDesign software (settings: Tier 3; Boosting: Balanced; Probe Precedence: Reuse Existing) and synthesized by Agilent Technologies (Santa Clara, CA, United States). The set of RNA baits sequences used in the current Target Enrichment protocol is available at https://doi.org/10.5281/zenodo.8043219.

2.5 SureSelect$^{XT HS}$ targeted whole-genome sequencing of *Francisella tularensis* directly from biological samples

In parallel to culture attempts, we used the designed RNA oligonucleotide baits in order to capture and sequence *F. tularensis* genomes directly from the biological samples, similarly to previous studies (Pinto et al., 2016; Borges et al., 2021; Macedo et al., 2023). After DNA extraction (described above), whole-genome capture and sequencing of *Francisella tularensis* was performed following SureSelect XT HS Target Enrichment System for Illumina Multiplexed Sequencing protocol (G9702-90000, Version E0, November 2020, Agilent Technologies, Santa Clara, CA, United States) using the custom baits library described above, in a 1:5 dilution. Library preparation started with the fragmentation of high quality DNA using the Agilent's SureSelectXT HS Low input Enzymatic Fragmentation kit ("Method 2: Enzymatic DNA Fragmentation" option described in "Step 2."), and was further carried out following the manufacturer's instructions. Libraries fragments size, concentration and molarity were determined on a Fragment Analyzer system (Agilent, Santa Clara, CA, United States). Libraries were then sequenced on an Illumina MiSeq or NextSeq apparatus (Illumina, San Diego, CA, United States) (Supplementary Table 1). To evaluate the success and feasibility of the target capture and enrichment (TCE) on a routine surveillance basis (in terms of both time and costs), a single library was prepared from each sample and sequenced only once (despite the potential gains of resequencing). The sequencing reads (only reads mapping against *F. tularensis* strain FTNF002-00 genome) generated in the present study were deposited in the European Nucleotide Archive (ENA) (BioProject PRJEB63267). Detailed ENA accession numbers are described in Table 1.

2.6 Sequence data analysis

To assess the efficiency of the whole-genome capture and enrichment, we determined the percentage of reads that corresponded to *F. tularensis* in each sample (percentage of reads "on-target") by mapping the reads (before and after quality improvement) against the reference genome *Francisella tularensis* subsp. *holarctica* FTNF002-00 (acc. no. NC_009749.1) using bowtie2 v.2.4.2 (Langmead and Salzberg, 2012). The percentage of the reference genome covered and respective mean depth of coverage were evaluated using the *getCoverage* tool.[1]

The *de novo* assembly of the genomes was performed with INNUca v.4.2.2,[2] which, after reads' quality analysis (FastQC v0.11.5)[3] and cleaning (Trimmomatic v0.36) (Bolger et al., 2014), performs *de novo* assembly with SPAdes v3.14.0 (Bankevich et al., 2012) and post-assembly optimization with Pilon v1.23 (Walker et al., 2014). The reads and contigs were taxonomically classified using Kraken 2 v2.0.7-beta (Wood et al., 2019) with the Standard-16 16 GB database (26th September 2022, available at: https://benlangmead.github.io/aws-indexes/k2). For the samples with a flag in the INNUca pipeline indicating that other taxa

[1] https://github.com/monsanto-pinheiro/getCoverage

[2] https://github.com/B-UMMI/INNUca

[3] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

TABLE 1 Details of the 17 *Francisella tularensis* genome assemblies generated in this study.

									Isolates metadata		
Sample ID	Approach	Assembly size (bp)	Assembly mean depth of coverage	No. Contigs (final assembly)	No. excluded contigs	N50	Genetic clade*	Acc. no. raw reads**	Location	Collection date	Host
FT_L09	$SS^{XT\ HS}$	1,764,191	442.1	138	0	20,843	B.49†	ERR11573837	Valladolid	2016	*L. granatensis*
FT_L10	$SS^{XT\ HS}$	1,766,793	415.4	128	0	22,699	B.49†	ERR11573838	Valladolid	2016	*L. granatensis*
FT_L11	$SS^{XT\ HS}$	1,760,164	426.0	125	0	23,357	B.108	ERR11573839	Valladolid	2016	*L. granatensis*
FT_L36	$SS^{XT\ HS}$	1,755,295	477.4	126	0	22,725	B.49	ERR11573840	Valladolid	2016	*L. granatensis*
FT_L51	$SS^{XT\ HS}$	1,773,693	475.7	130	0	22,725	B.153	ERR11573841	Palencia	2017	*L. granatensis*
FT_L52	$SS^{XT\ HS}$	1,775,015	577.4	125	0	22,726	B.266	ERR11573842	Palencia	2017	*L. granatensis*
FT_L71B	$SS^{XT\ HS}$	1,763,718	308.1	132	0	22,699	B.108	ERR11573843	Palencia	2019	*L. granatensis*
FT_L72B	$SS^{XT\ HS}$	1,774,977	357.1	153	0	20,929	B.153	ERR11573844	Palencia	2019	*L. granatensis*
FT_MA1830	$SS^{XT\ HS}$	1,780,847	1316.3	114	0	26,515	B.110	ERR11573845	Palencia	2014	*M. arvalis*
FT_MA1992	$SS^{XT\ HS}$	1,762,592	226.7	385	1,149	8,822	B.110	ERR11573846	Palencia	2014	*M. arvalis*
FT_MA2111	$SS^{XT\ HS}$	1,725,569	319.1	242	182	12,196	B.110	ERR11573847	Palencia	2014	*M. arvalis*
FT_MA2129	$SS^{XT\ HS}$	1,776,150	1117.4	114	0	25,220	B.110	ERR11573848	Palencia	2014	*M. arvalis*
FT_MA2136	$SS^{XT\ HS}$	1,718,841	104.7	274	277	10,565	B.110	ERR11573849	Palencia	2014	*M. arvalis*
FT_MA2129-Lung	Culture	1,789,044	147.4	103	NA	26,622	B.110	ERR11573850	Palencia	2014	*M. arvalis*
FT_MA2129-Spleen	Culture	1,788,868	142.6	101	NA	26,986	B.110	ERR11573851	Palencia	2014	*M. arvalis*
FT_MA1953-Liver	Culture	1,788,790	158.3	101	NA	26,986	B.110	ERR11573852	Palencia	2014	*M. arvalis*
FT_MA1953-Spleen	Culture	1,788,964	108.0	102	NA	26,622	B.110	ERR11573853	Palencia	2014	*M. arvalis*

$SS^{XT\ HS}$ – SureSelect $^{XT\ HS}$ target and enrichment protocol. NA, Not applicable.

*Genetic clade determined by CanSNPer2.

**Published reads of samples processed with $SS^{XT\ HS}$ correspond only to raw reads mapping against the reference genome *Francisella tularensis* subsp. *holarctica* FTNF002-00 (NC_009749.1).

†Samples FT_L09 and FT_L10 have a mixed infection with two genotypes – clade B.49 (major population) and B.110 (minor population). See main text and Supplementary Figure 7.

besides *Francisella* were identified in the assembly, only the contigs classified within the family *Francisellaceae*, or at lower taxonomical ranks, were kept in the final assemblies. When needed, genome annotation was performed with Prokka v1.14.6 (Seemann, 2014). The de novo assemblies generated in this study are available at https://doi.org/10.5281/zenodo.8043219.

2.7 Genetic diversity of the captured genomes and phylogenetic analysis

The genomes were genotyped using CanSNPer2,[4] which assigns the genomes to genetic clades based on the identification of established canonical single nucleotide polymorphisms (canSNPs) (Lärkeryd et al., 2014). To assess the phylogenetic context of the newly sequenced genomes into the *F. tularensis* diversity, all the *F. tularensis* complete genomes available in RefSeq and Genbank (and associated metadata), as of 27th February 2023, were collected and also classified with CanSNPer2. The details of the genomes can be found in Supplementary Table 2, including BioProjects and associated publications (La Scola et al., 2008; Barabote et al., 2009; Antwerpen et al., 2013; Coolen et al., 2013; Alm et al., 2015; Atkins et al., 2015; Dwibedi et al., 2016; Madani et al., 2017; Busch et al., 2018; Koene et al., 2019; Sichtig et al., 2019; Busch et al., 2020; Kevin et al., 2020; Kittl et al., 2020; Myrtennäs et al., 2020; Witt et al., 2020; Neubert et al., 2021; Öhrman et al., 2021; Pisano et al., 2021).

Multi-genome alignments and extraction of single nucleotide variant sites (SNVs) was performed with Parsnp v1.2 (Treangen et al., 2014) with the parameters *–c* and *–C 2000*, using the genome of strain FTNF002-00 (acc. no. NC_009749.1) as reference (for increased phylogenetic resolution, other reference genomes were used for some subclades phylogenetic trees; see figures legends for further information). Maximum likelihood trees were obtained with MEGA-CC v10.0.5 (Kumar et al., 2018) with 100 bootstraps. For particular comparative analyses, Snippy v4.5.1[5] was used for mapping and variant calling directly from the sequencing reads (with default parameters, with exception of *–mapqual* that was set to 20) and lofreq[6] was run over snippy BAM files for the detection of minor variants (Wilm et al., 2012), with *indelqual* mode with *–dindel* to assess indel qualities and then *call* mode, including *–call-indels*. Only the minor variants at a frequency of ≥5%, supported by at least 4 reads in positions covered by at least 10 reads were considered and further inspected with the Integrative Genomics Viewer (IGV) for confirmation/exclusion (Thorvaldsdottir et al., 2013).

Grapetree[7] (Zhou et al., 2018) and Microreact[8] (Argimón et al., 2016) were used for phylogenies and metadata visualization.

3 Results

3.1 Dataset characterization and culture of samples

A total of 63 samples, including samples from 50 common voles and 13 Iberian hares, that consisted of pooled of spleen and liver homogenates, were subjected to the SureSelect$^{XT\ HS}$ target capture and enrichment protocol (TCE) after DNA extraction. The counts and concentration of genome equivalents (GE) were assessed for 52 samples (46 voles and 6 hares) based on the screening of *tul4* and *ISFtu2* (Supplementary Table 1). All samples tested positive for *ISFtu2*, but 11 were negative for *tul4* in the real time PCR, but positive by conventional PCR followed by RLB. This discrepancy is likely due to the fact that ISFtu2 is an insertion element-like sequence present in multiple copies and hence its detection is more sensitive than *tul4* (Versage et al., 2003). All samples were confirmed as *F. tularensis subsp. holarctica*.

In parallel, culture was attempted on different tissues collected from the 50 voles (Supplementary Table 1). It was possible to isolate *F. tularensis* in two animals (4%), specifically from the spleen and liver of vole FT-MA1953 and from the lung and spleen of vole FT-MA2129. Concordantly, FT-MA1953 and FT-MA2129 were the voles samples with the highest GE counts used for TCE input (Supplementary Table 1). The four isolates were subjected to WGS and included in the downstream analysis.

3.2 Target capture and enrichment success

All 63 samples met the minimum required DNA input (10 ng) for the TCE protocol, with the majority having the maximum quantity of 200 ng of DNA available to use as input. After the TCE protocol, 14/63 samples were excluded due to very low library concentrations and the remaining 49 samples proceeded to NGS (Supplementary Table 1). A mean of 5,888,432 reads was generated per sample, with a high variability between samples (range: 169366–24,726,074 reads), although not correlating with success (Figure 1).

To evaluate the success of the TCE approach, the proportion of reads corresponding to *F. tularensis* (reads on-target) and the completeness of the genome (percentage of the reference genome covered by at least 1-fold) was determined for each sample. Among the 49 samples sequenced, 13 (26.5%) samples yielded >50% reads on-target (10 of which having >90% of reads on-target) and > 94% of the reference genome covered (Figure 1). Most of the remaining 36 unsuccessful samples had very low percentages of reads on-target (ranging between 0.01 and 10.64%, mean = 0.61%) and a mean of 9.3% of the genome covered (Figure 1). Interestingly, this revealed an "all-or-nothing" scenario where the successful and unsuccessful samples are clearly separated in two distinct groups (Figure 1B), rather than showing a linear trend of the evaluated metrics. In fact, success was obtained for all samples with more than 1×10^5 GE of input, with a minimum of 463,513 GE for sample FT-MA2136, while the highest GE input among the "fail" group was 13,948 GE (Figure 1B). This large "gap" in GE input values between success and fail groups might help explain the "all-or-nothing" scenario with no samples with intermediate success in between.

[4] https://github.com/FOI-Bioinformatics/CanSNPer2
[5] https://github.com/tseemann/snippy
[6] https://github.com/CSB5/lofreq
[7] https://github.com/achtman-lab/GrapeTree
[8] https://microreact.org/

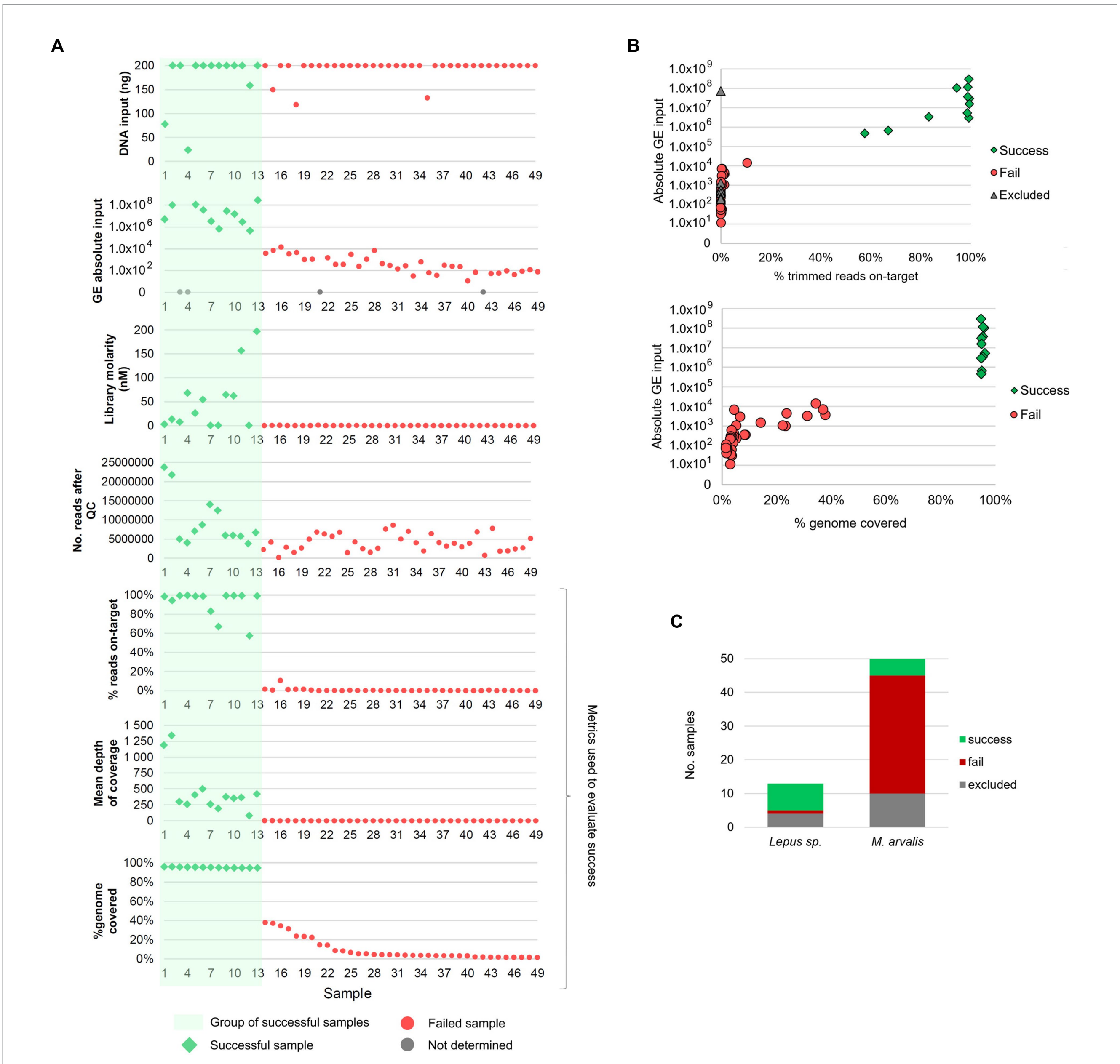


FIGURE 1
Overview of the success of the target and enrichment sequencing approach and its relation to multiple sample features. **(A)** Relation between genome coverage and percentage of reads on-target (three bottom panels), and multiple factors: initial DNA input, absolute input of genome equivalents (GE), NGS library molarity and number of reads after quality control (QC). Samples are ordered from 1 to 49 (on the *xx*-axis) from the sample with the highest to the lowest percentage of horizontal genome coverage (by at least 1-fold). **(B)** Association between absolute GE input [available for 52 samples (45 sequenced and 7 excluded)] and: (i) the percentage of reads on-target (upper panel) and, (ii) the percentage of genome coverage (lower panel). Gray triangles in the upper panel identifies the samples that were excluded before sequencing due to low library concentration but for which GE counts were available (*n*=7). The excluded sample with high GE counts (1×10^8) corresponds to sample FT-MA1953. **(C)** Distribution of successful, failed and excluded samples among the samples from common voles (*M. arvalis*, *n*=50) and Iberian hares (*L. granatensis*, *n*=13).

Among the 50 vole samples that were also subjected to culture, isolation of *F. tularensis* was possible for two samples (4%), while the culture-free TCE approach allowed obtaining the genome of *F. tularensis* from five samples (10%). Of note, one of the two successfully cultured samples, FT-MA1953, did not proceed to NGS after the TCE protocol due to a very low library concentration. However, the high GE count (~7×10^7) of FT-MA1953 falls within the successful GE input range and it is, in fact, the only sample failing within this group (Figure 1B). As such, we speculate that some experimental issue might have occurred during the wet-lab protocol leading to a poor library (e.g., failing to add a reagent) for this sample.

Of note, TCE success rate was much higher among hares (61.5%) than voles (10%) samples (Figure 1C), consistent with the higher GE counts observed in the former species (Supplementary Table 1).

3.3 Characterization of *Francisella tularensis* genomes and integration into the global phylogeny

3.3.1 CanSNP classification

The *de novo* assemblies of the 13 successful sequenced genomes ranged from 1.72 to 1.78 Mb, which is close to the expected *F. tularensis* genome size (1.8–1.9 Mb), and had a mean depth of coverage of 505-fold (range 104–1,316-fold) (Table 1). Genomes were typed based on canSNPs classification by canSNPer2 (Lärkeryd et al., 2014). For clarity reasons, the clades are referred to according to their level of discrimination (from D0 to D9) within this dataset (Figure 2), where level D0 corresponds to the level of the clade that is common to all the analyzed sequences and level D1 is the first level where the analyzed sequences start being discriminated by clade.

All 13 genomes belonged to major clade B.6 (level D0) (descendant of ancestral clades B.1-B.2-B.3-B.5), consistent with strains circulating in Europe and North America (Pilo, 2018). Among these, 11 genomes belonged to subclade B.10 (D1), common in Western Europe, and the

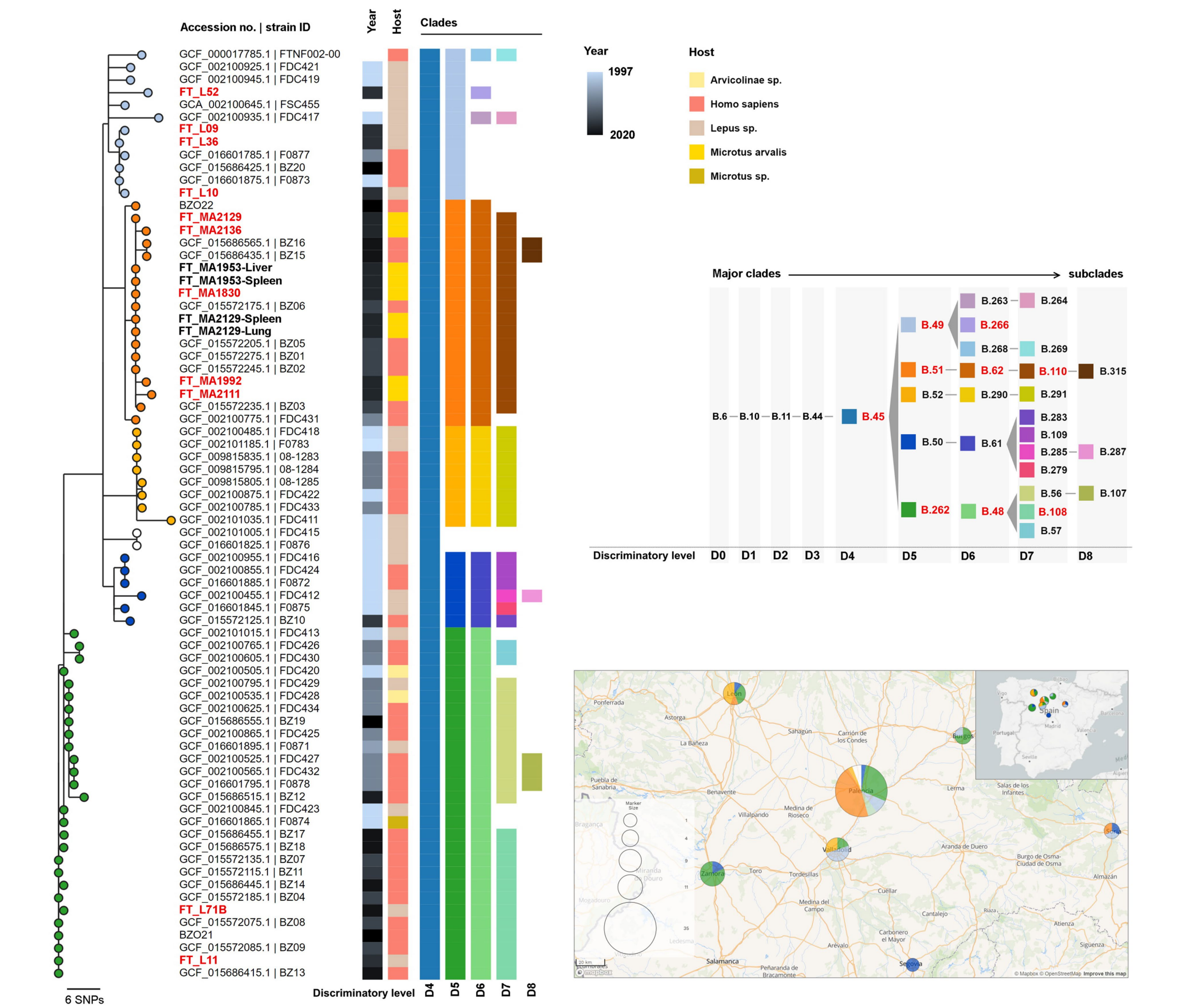


FIGURE 2
Maximum likelihood phylogenetic tree of the newly sequenced genomes from clade B.10 (D1; *n* = 17) and all available *Francisella tularensis subsp. holarctica* genome assemblies from Spain (*n* = 58; Supplementary Table 2). The phylogenetic tree was generated using MEGA-CC v10.0.5 with 100 bootstraps based on 113 core single nucleotide variant positions (SNVs) extracted from a multiple genome alignment with 1,558,560 bp, generated with parsnp using the genome of strain FTNF002-00 (acc. no. NC_009749.1) as reference. Tree nodes are colored by clade (at level D5) and metadata blocks show the respective year of collection, host and canSNP clades and subclades (from level D4 to D8). The strain IDs from samples generated in this study using the capture and enrichment approach (TCE) (*n* = 11) are shown in red. Other genomes obtained from strains isolated in culture are shown in black and bold (*n* = 4). The panel on the right shows the color codes for each clade and subclade identified among the analyzed genomes and the ancestry relationships between them. The clades identified among the TCE genomes are highlighted in red. The ancestral clades (SNP path) of major clade B.45 identified by CanSNPer2 are B.1-B.2-B.3-B.5-B.6-B.10-B.11-B.44-B.45 (not shown). The lower panel on the right shows the geographical distribution of the samples, with the same sample color scheme of the nodes on the phylogenetic tree (clade level D5). A smaller map shows the geographical location of the study area within the Iberian Peninsula.

other two belonged to B.7 (D1), a common clade in the United States and Scandinavia (Supplementary Figure 1) (Pilo, 2018).

3.3.2 Integration in global phylogeny

All the genome assemblies of *F. tularensis* subs. *holartica* (or *F. tularensis* without subspecies described) obtained from public databases (RefSeq and Genbank) were classified with CanSNPer2. All the non-duplicate genomes from subclades B.10 and B.7 (level D1) were selected for the phylogenetic analysis (*n* = 556, Supplementary Table 2, Supplementary Figure 1).

The 11 B.10 (level D1) genomes belonged to subclade B.45 (level D4) and could be further discriminated into three subclades within level D5: (i) B.49 (*n* = 4), (ii) B.51 (*n* = 5) and (iii) B.262 (*n* = 2) (Figure 2 and Supplementary Figure 2, Table 1).

Within the global phylogeny of subclade B.49 (*n* = 81) (Supplementary Figure 3), which included mostly genomes from France (n = 45), Germany (*n* = 21) and Spain (*n* = 11), the genomes from samples FT_L09, FT_L10, and FT_L36 segregated in a branch together with other strains from Spain (three human isolates) and one from France (isolated from hare) (Figure 2 and Supplementary Figure 3). Strain FT_L52 integrates another phylogenetic branch that is composed of six B.266 (level D6) genomes, four of which are from France (including one human isolate).

Within B.51 (level D5), five genomes obtained through the TCE approach (FT_MA1830, FT_MA1992, FT_MA2111, FT_MA2129, and FT_MA2136) belonged to B.62 (level D6) and B.110 (level D7) subclades. The B.1.110 subclade (total of 19 sequences) also included the four isolates obtained from culture (FT_MA1953-Liver, FT_MA1953-Spleen, FT_MA2129-Lung, and FT_MA2129-Spleen), seven human isolates from Spain and three hare isolates from Germany (Supplementary Figure 4). Interestingly, although B.110 is strongly associated with samples collected in Spain, its ascendant clade B.51 (level D5; *n* = 59) is mostly associated with genomes from Germany (Supplementary Figure 4). Furthermore, it is noteworthy that all the B.110 strains are very closely related, even when comparing strains from Spain and Germany.

Within B.262 (level D5), both studied samples (FT_L11 and FT_L71B) belonged to subclade B.108 (D7), which involves nine other human isolates from Spain (Figure 2 and Supplementary Figure 5). This subclade is a descendent of B.48 (level D6), which is almost exclusively associated with samples collected in Spain (Supplementary Figure 5), concordant with previous reports (Dwibedi et al., 2016).

The two B.7 (level D1) genomes obtained in this study, FT_L51 and FT_L72B [further classified in subclades B.133 (D2), B.81 (D3), B.136 (D4), and B.153 (D5)], were analyzed together with 26 other genome assemblies available for this major clade (Supplementary Figure 1). These genomes originate from Norway, Germany, United States, Sweden, Finland and Russia, making FT_L51 and FT_L72B the only samples from Spain among this clade. Supplementary Figure 6 shows the phylogenetic relationship of the 22 closest isolates (USA genomes are not shown as they cluster apart at a high genetic distance).

3.3.3 Comparing the genomic information obtained by TCE and culture

Among the studied sample dataset, we could obtain the genome of *F. tularensis* from both culture and TCE approaches for vole FT_MA2129. No phylogenetic differences were observed between the three genomes (FT_MA2129 obtained by TCE and the culture isolates obtained from the lung FT_MA2129-Lung and spleen FT_MA2129-Spleen) obtained from this vole (Figure 2 and Supplementary Figure 4), which could be confirmed by further fine-tuned mapping-based SNP calling. Furthermore, the TCE-derived *de novo* assembly of sample FT_MA2129 was only about ~12 kb shorter (with similar number of contigs and N50) than the assemblies generated from the cultured isolates (Table 1). Although no TCE genome was obtained from vole FT_MA1953, a mapping-based SNP analysis was also applied to compare the two culture genomes from this animal (FT_MA1953-Liver and FT_MA1953-Spleen), confirming that they had no differences.

To further access the genomic information obtained with TCE, we explored the minor variants (i.e., low frequency populations) in sample FT_MA2129. For this, we mapped the reads of all three samples FT_MA2129, FT_MA2129-Spleen, and FT_MA2129-Lung against the FT_MA2129-Lung genome assembly and detected the minor variants with ≥5% frequency. We validated 13 minor variants in sample FT_MA2129 (with frequencies ranging between 5.1% and 13.8%) that were not detected in either of the two culture-derived genomes (which had only one minor variant validated each) (Supplementary Table 3). The proteins predicted to be affected by the 13 mutations include several transporters and two proteins potentially involved in host-pathogen interaction (ankyrin repeat domain-containing protein and normocyte binding protein 2b), among others (Supplementary Table 3).

3.3.4 Detection of mixed infections

The screening of minor variants was also performed for the remaining samples and revealed multiple mutations at frequencies between 20 and 50% in samples FT_L09 and FT_L10. To assess if these mutations corresponded to subpopulations of the same strain or to a different strain in the same sample, we artificially inserted the minor mutations in the respective original assemblies. The CanSNPer2 classification of these assemblies revealed that the minor populations belonged to clades B.110 (D7) (descendant of B.51 (D5) and B.62 (D6)) contrarily to the major populations in these samples that belong to genotype B.49 (D5), confirming that the two hares had mixed infections with two different genotypes. Moreover, the integration of these genomes (FT_L09_MINOR and FT_L10_MINOR) in the phylogenetic analysis of the dataset from Spain show that they cluster perfectly within clade B.110 (Supplementary Figure 7), together with the other five TCE genomes already in this clade.

4 Discussion

In this study, we were able to sequence near-complete *F. tularensis* genomes using a culture-free approach relying on the use of RNA oligonucleotide "baits" to capture and enrich *F. tularensis* genomic material directly from complex biological samples. The high-quality sequencing data generated allowed the characterization and typing of the newly sequenced genomes and their integration into the global phylogeny of *F. tularensis subsp. holarctica*. The new genomes belonged to clades B.10 (D1), common in Western Europe, and unexpectedly to B.7 (D1), common in the United States and Scandinavia (Pilo, 2018).

The TCE approach here described had a higher success rate than the conventional culture-based method. A recent study by Wagner et al. (2022) has applied a similar approach with Agilent SureSelect technology (Agilent Technologies, Santa Clara, CA, United States) where RNA probes were designed to detect and characterize the genome of *F. tularensis* and other diverse species in the family *Francisellaceae*. Although a high success among samples with low concentration of target DNA (including environmental samples) was reported, there was no information regarding a direct or indirect quantification of the pathogen in the samples, which hampers a direct comparison with the success rate achieved in the present work. Furthermore, Wagner and colleagues processed the samples with two rounds of TCE, considerably increasing processing time and cost (Wagner et al., 2022). Still, as a potential gain in sensitivity may be achieved, the TCE two-round strategy applied by Wagner and colleagues may be a very interesting option. Our approach had an estimated cost of ~250€ per sample (for SureSelect and Illumina reagents only) and a turnaround time of four to five working days (from sample processing to sequence data), presenting a valuable option to recover near-complete genomic information.

The implementation of a TCE protocol also largely benefits from a pre-selection of samples with higher probability of success. Here, we showed that samples with at least 1×10^5 GE of input had a success rate close to 100%. As such, the GE concentration estimated from real-time PCR data provides a straightforward approach for sample pre-selection using this "cut-off" value, increasing the effectiveness of the protocol.

Of note, we observed a higher success rate of TCE sequencing of hares samples (which had much higher GE concentrations) compared to voles samples. One explanation for this might be the capture context of these animals since, unlike voles (that were captured with traps), some of the sampled hares might have died of tularemia (7/13 were found dead) and others were hunted or hit by car, hinting that they might have been debilitated. Thus, in a pure speculative basis, if these animals died (or were sick) with tularemia, it is likely that they had much higher bacterial loads (congruent with our observations).

Comparing the sequence data generated from the same sample with both TCE and culture-based methods (sample FT_MA2129) showed that the designed RNA baits are highly efficient to capture the full genomic information of *F. tularensis* and that the culture-free generated data is highly reliable and equivalent to that obtained by sequencing of the isolates. This observation indicates that the 13 newly sequenced genomes could be well characterized at phylogenetic level. Importantly, CanSNP-based typing showed that the new genomes belonged to different clades commonly identified in Europe and was highly congruent with the genome-based phylogenies. Further phylogenetic analysis indicated that most of the new genomes were closely related to other genomes from Spain (Figure 2), with the noteworthy exception of the two new genomes from clade B.7, which is more commonly found in Scandinavia and the United States (Supplementary Figure 6). As previously reported, and in accordance with the known low genetic diversity of the subspecies, subclades do not generally correlate with geography or host (Dwibedi et al., 2016; Kevin et al., 2020), as is well illustrated here by the identification of several closely related strains largely dispersed in different countries and different hosts.

Interestingly, the three main clades (at the D5 discriminatory level) found in Spain, B.49, B.51, and B.262 (Supplementary Figure 2), were identified in both human and hares (B.49 and B.52; Figure 2) during the initial outbreaks (1997–2007) and later identified in humans and common voles (B.51; Figure 2) during the most recent vole outbreaks (2014 onwards). As such, our findings are consistent with the important roles of hares and voles outbreaks in the tularemia epidemiology and transmission to humans in this region (e.g., 2014 outbreak; see Ariza-Miguel et al., 2014; Luque-Larena et al., 2015; Herrero-Cófreces et al., 2021). Remarkably, using TCE sequencing, in two hares we were able to identify mixed infections with two distinct genotypes (clades B.49 and B.51), which further highlights the role of this host species in the transmission dynamics of tularemia.

Furthermore, we could also identify minor variants in a sample subjected to TCE that were not present in the genomes obtained from the respective cultured isolates. We hypothesize that these subpopulations reflect the within-host pathogen diversity, potentially linked to the ongoing adaptation to different niches/tissues. While RNA "baits" acted upon complex biological samples (in this case, a homogenate of liver and spleen), potentially capturing the *in vivo* genomic diversity of the strain, sequencing from culture reflects only the diversity of one or a few colonies. This approach has been previously applied for other pathogens, such as *Treponema pallidum*, revealing substantial within-patient genetic diversity (Pinto et al., 2016), *Mycobacterium tuberculosis*, as an alternative approach for surveillance and drug susceptibility inferences (Macedo et al., 2023), and *Chlamydia trachomatis*, where the technique was used to characterize and monitor the transcontinental dissemination of an emergent recombinant strain (Borges et al., 2021). In this context, despite *F. tularensis* being a highly monomorphic pathogen, the TCE approach allows the exploration of this additional layer of genetic variability of low-frequency populations and potentially distinguish isolates that are otherwise indistinguishable (if only the consensus sequences are considered).

Globally, this study showed that the TCE method can generate complete *F. tularensis* genomic information in a timely manner, allowing us to carry out highly discriminatory phylogenetic analysis at whole-genome level. We performed CanSNP typing, integration of genomes into global phylogeny (with important observations in regards to clades geographical distribution and hosts) and fine-tuned mutation analysis, showing that TCE can greatly increase the current knowledge on *F. tularensis*. Although the applicability of target capture and enrichment for sequencing on a routine surveillance basis is still debatable (complexity of protocols, cost), TCE performed with greater success than culture-dependent sequencing and can be further potentiated by the pre-selection of samples based on real-time PCR data, as we propose. Considering the low success rate of *F. tularensis* culture and the fact that serology, a main diagnosis method, provides no information regarding molecular epidemiology, the present methodology provides a highly valuable approach toward an increased knowledge on the genomics and epidemiology of this highly infectious pathogen.

Data availability statement

Sequencing reads (only reads mapping against F. tularensis strain FTNF002-00 genome) generated in the present study were deposited in the European Nucleotide Archive (ENA) (BioProject PRJEB63267). Detailed ENA accession numbers are described in Table 1. The set of

RNA baits sequences used in the current Target Enrichment protocol and the de novo assemblies are available at https://doi.org/10.5281/zenodo.8043219.

Ethics statement

The animal study was approved by the Ethics Committee from the University of Valladolid (authorization code: 4801646). The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

JI: Investigation, Methodology, Writing – original draft, Data curation, Formal analysis, Validation. RE: Investigation, Methodology, Writing – original draft, Conceptualization, Funding acquisition, Project administration, Resources, Supervision. JL-L: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – review & editing. MP: Investigation, Methodology, Writing – review & editing, Data curation. VB: Data curation, Investigation, Methodology, Writing – review & editing, Validation. RG-M-N: Investigation, Writing – review & editing, Resources. SD: Writing – review & editing, Data curation, Methodology. LV: Data curation, Methodology, Writing – review & editing. FM: Methodology, Writing – review & editing, Conceptualization, Funding acquisition, Investigation, Resources, Supervision. DV: Conceptualization, Investigation, Methodology, Resources, Supervision, Writing – review & editing. DH-R: Investigation, Resources, Writing – review & editing, Data curation, Visualization. RR-P: Data curation, Investigation, Resources, Writing – review & editing. SH-C: Data curation, Investigation, Resources, Writing – review & editing. FJ-T: Resources, Writing – review & editing. JG: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft. IL: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by funding from the European Union EU4Health Programme under grant agreement No 101113460 (GENEO) and by the GenomePT project (POCI-01-0145-FEDER-022184), supported by COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by Fundação para a Ciência e a Tecnologia (FCT). The acquisition of WGS-related reagents was supported by the Ministerio de Economía y Competitividad, Government of Spain/FEDER, RESERTULA project (CLG2015-66962-C2-2-R). Fieldwork for sample collection was funded by the Ministerio de Economía y Competitividad, Government of Spain/FEDER ECOTULA project (CGL2015-66962-C2-1-R) and by the Ministerio de Ciencia, Innovación y Universidades BOOMRAT project (PID2019-109327RB-I00) from Spain.

Acknowledgments

The authors want to thank Carolina Nunes for the support in laboratory work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1277468/full#supplementary-material

References

Alm, E., Advani, A., Bråve, A., and Wahab, T. (2015). Draft genome sequence of strain R13-38 from a *Francisella tularensis* outbreak in Sweden. *Genome Announc.* 3:e01517–14. doi: 10.1128/genomeA.01517-14

Aloni-Grinstein, R., Schuster, O., Yitzhaki, S., Aftalion, M., Maoz, S., Steinberger-Levy, I., et al. (2017). Isolation of Francisella tularensis and Yersinia pestis from blood cultures by plasma purification and Immunomagnetic separation accelerates antibiotic susceptibility determination. *Front. Microbiol.* 8:312. doi: 10.3389/fmicb.2017.00312

Antwerpen, M. H., Schacht, E., Kaysser, P., and Splettstoesser, W. D. (2013). Complete genome sequence of a *Francisella tularensis* subsp. holarctica strain from Germany causing lethal infection in common marmosets. *Genome Announc.* 1:e00135–12. doi: 10.1128/genomeA.00135-12

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., et al. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.* 2:e000093. doi: 10.1099/mgen.0.000093

Ariza-Miguel, J., Johansson, A., Fernández-Natal, M. I., Martínez-Nistal, C., Orduña, A., Rodríguez-Ferri, E. F., et al. (2014). Molecular investigation of tularemia outbreaks, Spain, 1997–2008. *Emerg. Infect. Dis.* 20, 754–761. doi: 10.3201/eid2005.130654

Atkins, L. M., Holder, M. E., Ajami, N. J., Metcalf, G. A., Weissenberger, G. M., Wang, M., et al. (2015). High-quality draft genome sequence of Francisella tularensis subsp. holarctica strain OR96-0246. *Genome Announc.* 3:e00898–15. doi: 10.1128/genomeA.00898-15

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Barabote, R. D., Xie, G., Brettin, T. S., Hinrichs, S. H., Fey, P. D., Jay, J. J., et al. (2009). Complete genome sequence of Francisella tularensis subspecies holarctica FTNF002-00. *PLoS One* 4:e7041. doi: 10.1371/journal.pone.0007041

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

Borges, V., Isidro, J., Correia, C., Cordeiro, D., Vieira, L., Lodhia, Z., et al. (2021). Transcontinental dissemination of the L2b/D-Da recombinant chlamydia trachomatis lymphogranuloma venereum (LGV) strain: need of broad multi-country molecular surveillance. *Clin. Infect. Dis.* 73, e1004–e1007. doi: 10.1093/cid/ciab067

Busch, A., Homeier-Bachmann, T., Abdel-Glil, M. Y., Hackbart, A., Hotzel, H., and Tomaso, H. (2020). Using affinity propagation clustering for identifying bacterial clades and subclades with whole-genome sequences of *Francisella tularensis*. *PLoS Negl. Trop. Dis.* 14:e0008018. doi: 10.1371/journal.pntd.0008018

Busch, A., Thomas, P., Zuchantke, E., Brendebach, H., Neubert, K., Gruetzke, J., et al. (2018). Revisiting Francisella tularensis subsp. holarctica, causative agent of tularemia in Germany with bioinformatics: new insights in genome structure, DNA methylation and comparative phylogenetic analysis. *Front. Microbiol.* 9:344. doi: 10.3389/fmicb.2018.00344

Carvalho, C. L., Lopes de Carvalho, I., Zé-Zé, L., Núncio, M. S., and Duarte, E. L. (2014). Tularaemia: A challenging zoonosis. *Comp. Immunol. Microbiol. Infect. Dis.* 37, 85–96. doi: 10.1016/j.cimid.2014.01.002

Centers for Disease Control and Prevention (CDC) (n.d.). Bioterrorism agents/diseases. Available at: https://emergency.cdc.gov/agent/agentlist-category.asp#a (accessed July 3, 2023).

Coolen, J. P. M., Sjödin, A., Maraha, B., Hajer, G. F., Forsman, M., Verspui, E., et al. (2013). Draft genome sequence of *Francisella tularensis* subsp. holarctica BD11-00177. *Stand. Genomic Sci.* 8, 539–547. doi: 10.4056/sigs.4217923

Dennis, D. T., Inglesby, T. V., Henderson, D. A., Bartlett, J. G., Ascher, M. S., Eitzen, E., et al. (2001). Tularemia as a biological weapon: medical and public health management. *JAMA* 285, 2763–2773. doi: 10.1001/jama.285.21.2763

Doern, G. V. (2000). Detection of selected fastidious Bacteria. *Clin. Infect. Dis.* 30, 166–173. doi: 10.1086/313586

Dwibedi, C., Birdsell, D., Lärkeryd, A., Myrtennäs, K., Öhrman, C., Nilsson, E., et al. (2016). Long-range dispersal moved Francisella tularensis into Western Europe from the east. *Microb. Genom.* 2:e000100. doi: 10.1099/mgen.0.000100

Escudero, R., Toledo, A., Gil, H., Kovácsová, K., Rodríguez-Vargas, M., Jado, I., et al. (2008). Molecular method for discrimination between Francisella tularensis and Francisella -like endosymbionts. *J. Clin. Microbiol.* 46, 3139–3143. doi: 10.1128/JCM.00275-08

Herrero-Cófreces, S., Mougeot, F., Lambin, X., and Luque-Larena, J. J. (2021). Linking zoonosis emergence to farmland invasion by fluctuating herbivores: common vole populations and tularemia outbreaks in NW Spain. *Front. Vet. Sci.* 8:454. doi: 10.3389/fvets.2021.698454

Johansson, A., and Petersen, J. M. (2010). *Genotyping of Francisella tularensis*, the causative agent of tularemia. *J. AOAC Int.* 93, 1930–1943. doi: 10.1093/jaoac/93.6.1930

Kevin, M., Girault, G., Caspar, Y., Cherfa, M. A., Mendy, C., Tomaso, H., et al. (2020). Phylogeography and genetic diversity of *Francisella tularensis* subsp. holarctica in France (1947–2018). *Front. Microbiol.* 11:286. doi: 10.3389/fmicb.2020.00287

Kittl, S., Francey, T., Brodard, I., Origgi, F. C., Borel, S., Ryser-Degiorgis, M.-P., et al. (2020). First European report of *Francisella tularensis* subsp. holarctica isolation from a domestic cat. *Vet. Res.* 51:109. doi: 10.1186/s13567-020-00834-5

Koene, M., Rijks, J., Maas, M., Ruuls, R., Engelsma, M., van Tulden, P., et al. (2019). Phylogeographic distribution of human and hare Francisella Tularensis Subsp. Holarctica strains in the Netherlands and its pathology in European Brown hares (Lepus Europaeus). *Front. Cell. Infect. Microbiol.* 9:11. doi: 10.3389/fcimb.2019.00011

Kugeler, K., Pappert, R., Zhou, Y., and Petersen, J. (2006). Real-time PCR for *Francisella tularensis* types a and B. *Emerg. Infect. Dis.* 12, 1799–1801. doi: 10.3201/eid1211.060629

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

La Scola, B., Elkarkouri, K., Li, W., Wahab, T., Fournous, G., Rolain, J.-M., et al. (2008). Rapid comparative genomic analysis for clinical microbiology: the *Francisella tularensis* paradigm. *Genome Res.* 18, 742–750. doi: 10.1101/gr.071266.107

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lärkeryd, A., Myrtennäs, K., Karlsson, E., Dwibedi, C. K., Forsman, M., Larsson, P., et al. (2014). CanSNPer: a hierarchical genotype classifier of clonal pathogens. *Bioinformatics* 30, 1762–1764. doi: 10.1093/bioinformatics/btu113

Luque-Larena, J. J., Mougeot, F., Arroyo, B., Vidal, M. D., Rodríguez-Pastor, R., Escudero, R., et al. (2017). Irruptive mammal host populations shape tularemia epidemiology. *PLoS Pathog.* 13:e1006622. doi: 10.1371/journal.ppat.1006622

Luque-Larena, J. J., Mougeot, F., Roig, D. V., Lambin, X., Rodríguez-Pastor, R., Rodríguez-Valín, E., et al. (2015). Tularemia outbreaks and common vole (*Microtus arvalis*) irruptive population dynamics in northwestern Spain, 1997–2014. *Vector-Borne Zoonotic Dis.* 15, 568–570. doi: 10.1089/vbz.2015.1770

Macedo, R., Isidro, J., Ferreira, R., Pinto, M., Borges, V., Duarte, S., et al. (2023). Molecular capture of *Mycobacterium tuberculosis* genomes directly from clinical samples: a potential backup approach for epidemiological and drug susceptibility inferences. *Int. J. Mol. Sci.* 24:2912. doi: 10.3390/ijms24032912

Madani, N., Giraud, P., Mendy, C., Colaneri, C., Cherchame, E., Cherfa, M.-A., et al. (2017). First draft genome sequences of three strains of *Francisella tularensis* subsp. holarctica, isolated from hares and a tick in France. *Genome Announc.* 5:e00993–17. doi: 10.1128/genomeA.00993-17

Myrtennäs, K., Escudero, R., Zaballos, Á., González-Martín-Niño, R., Gyuranecz, M., and Johansson, A. (2020). Genetic traces of the Francisella tularensis colonization of Spain, 1998–2020. *Microorganisms* 8:1784. doi: 10.3390/microorganisms8111784

Neubert, K., Zuchantke, E., Leidenfrost, R. M., Wünschiers, R., Grützke, J., Malorny, B., et al. (2021). Testing assembly strategies of *Francisella tularensis* genomes to infer an evolutionary conservation analysis of genomic structures. *BMC Genomics* 22:822. doi: 10.1186/s12864-021-08115-x

Öhrman, C., Sahl, J. W., Sjödin, A., Uneklint, I., Ballard, R., Karlsson, L., et al. (2021). Reorganized genomic taxonomy of Francisellaceae enables Design of Robust Environmental PCR assays for detection of *Francisella tularensis*. *Microorganisms* 9:146. doi: 10.3390/microorganisms9010146

Petersen, J. M., Mead, P. S., and Schriefer, M. E. (2009). *Francisella tularensis*: an arthropod-borne pathogen. *Vet. Res.* 40:07. doi: 10.1051/vetres:2008045

Petersen, J. M., Schriefer, M. E., Gage, K. L., Montenieri, J. A., Carter, L. G., Stanley, M., et al. (2004). Methods forenhanced culture recovery of *Francisella tularensis*. *Appl. Environ. Microbiol.* 70, 3733–3735. doi: 10.1128/AEM.70.6.3733-3735.2004

Pilo, P. (2018). Phylogenetic lineages of *Francisella tularensis* in animals. *Front. Cell. Infect. Microbiol.* 8:258. doi: 10.3389/fcimb.2018.00258

Pinto, M., Borges, V., Antelo, M., Pinheiro, M., Nunes, A., Azevedo, J., et al. (2016). Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat. Microbiol.* 2:16190. doi: 10.1038/nmicrobiol.2016.190

Pisano, S. R. R., Kittl, S., Eulenberger, U., Jores, J., and Origgi, F. C. (2021). Natural infection of a European red squirrel (*Sciurus vulgaris*) with *Francisella tularensis* subsp. *Holarctica. J. Wildl. Dis.* 57, 970–973. doi: 10.7589/JWD-D-20-00182

Rodríguez-Pastor, R., Escudero, R., Vidal, D., Mougeot, F., Arroyo, B., Lambin, X., et al. (2017). Density-dependent prevalence of *Francisella tularensis* in fluctuating vole populations, northwestern Spain. *Emerg. Infect. Dis.* 23, 1377–1379. doi: 10.3201/eid2308.161194

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

Shevtsov, V., Kairzhanova, A., Shevtsov, A., Shustov, A., Kalendar, R., Abdrakhmanov, S., et al. (2021). Genetic diversity of *Francisella tularensis* subsp. holarctica in Kazakhstan. *PLoS Negl. Trop. Dis.* 15:e0009419. doi: 10.1371/journal.pntd.0009419

Sichtig, H., Minogue, T., Yan, Y., Stefan, C., Hall, A., Tallon, L., et al. (2019). FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nat. Commun.* 10:3313. doi: 10.1038/s41467-019-11306-6

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

Versage, J. L., Severin, D. D. M., Chu, M. C., and Petersen, J. M. (2003). Development of a multitarget Real-time TaqMan PCR assay for enhanced detection of *Francisella tularensis* in complex specimens. *J. Clin. Microbiol.* 41, 5492–5499. doi: 10.1128/JCM.41.12.5492-5499.2003

Wagner, D. M., Birdsell, D. N., McDonough, R. F., Nottingham, R., Kocos, K., Celona, K., et al. (2022). Genomic characterization of Francisella tularensis and other diverse Francisella species from complex samples. *PLoS One* 17:e0273273. doi: 10.1371/journal.pone.0273273

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201. doi: 10.1093/nar/gks918

Witt, N., Andreotti, S., Busch, A., Neubert, K., Reinert, K., Tomaso, H., et al. (2020). Rapid and culture free identification of Francisella in hare carcasses by high-resolution tandem mass spectrometry Proteotyping. *Front. Microbiol.* 11:636. doi: 10.3389/fmicb.2020.00636

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biol.* 20:257. doi: 10.1186/s13059-019-1891-0

Zhou, Z., Alikhan, N.-F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., et al. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* 28, 1395–1404. doi: 10.1101/gr.232397.117

Frontiers in Genetics

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fgene.2023.1217255

Polyketide synthases mutation in tuberculosis transmission revealed by whole genomic sequence, China, 2011–2019

Ting-Ting Wang[1], Yuan-Long Hu[1†], Yi-Fan Li[2†], Xiang-Long Kong[3†], Ya-Meng Li[1], Ping-Yi Sun[4], Da-Xing Wang[5], Ying-Ying Li[1], Yu-Zhen Zhang[6], Qi-Lin Han[6], Xue-Han Zhu[6], Qi-Qi An[7], Li-Li Liu[5], Yao Liu[7]* and Huai-Chen Li[1,7]*

[1]Shandong University of Traditional Chinese Medicine, Jinan, China, [2]Department of Pulmonary and Critical Care Medicine, The Third Affiliated Hospital of Shandong First Medical University (Affiliated Hospital of Shandong Academy of Medical Sciences), Jinan, China, [3]Shandong Artificial Intelligence Institute Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, [4]Jining Medical University, Jining, China, [5]People's Hospital of Huaiyin Jinan, Jinan, China, [6]Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China, [7]Department of Respiratory and Critical Care Medicine, Shandong Provincial Hospital Affiliated to 11 Shandong University, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China

OPEN ACCESS

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Tulio de Lima Campos,
Aggeu Magalhães Institute (IAM), Brazil
Richard Salvato,
State Center for Health Surveillance,
Brazil

*CORRESPONDENCE
Huai-Chen Li,
✉ lihuaichen@163.com
Yao Liu,
✉ doctorliuyao@126.com

[†]These authors have contributed equally to this work

RECEIVED 20 May 2023
ACCEPTED 30 November 2023
PUBLISHED 08 January 2024

CITATION
Wang T-T, Hu Y-L, Li Y-F, Kong X-L, Li Y-M, Sun P-Y, Wang D-X, Li Y-Y, Zhang Y-Z, Han Q-L, Zhu X-H, An Q-Q, Liu L-L, Liu Y and Li H-C (2024), Polyketide synthases mutation in tuberculosis transmission revealed by whole genomic sequence, China, 2011–2019. *Front. Genet.* 14:1217255. doi: 10.3389/fgene.2023.1217255

Introduction: Tuberculosis (TB) is an infectious disease caused by a bacterium called *Mycobacterium tuberculosis* (*Mtb*). Previous studies have primarily focused on the transmissibility of multidrug-resistant (MDR) or extensively drug-resistant (XDR) *Mtb*. However, variations in virulence across *Mtb* lineages may also account for differences in transmissibility. In *Mtb*, polyketide synthase (PKS) genes encode large multifunctional proteins which have been shown to be major mycobacterial virulence factors. Therefore, this study aimed to identify the role of PKS mutations in TB transmission and assess its risk and characteristics.

Methods: Whole genome sequences (WGSs) data from 3,204 *Mtb* isolates was collected from 2011 to 2019 in China. Whole genome single nucleotide polymorphism (SNP) profiles were used for phylogenetic tree analysis. Putative transmission clusters (≤10 SNPs) were identified. To identify the role of PKS mutations in TB transmission, we compared SNPs in the PKS gene region between "clustered isolates" and "non-clustered isolates" in different lineages.

Results: Cluster-associated mutations in *ppsA, pks12,* and *pks13* were identified among different lineage isolates. They were statistically significant among clustered strains, indicating that they may enhance the transmissibility of *Mtb*.

Conclusion: Overall, this study provides new insights into the function of PKS and its localization in *M. tuberculosis*. The study found that ppsA, pks12, and pks13 may contribute to disease progression and higher transmission of certain strains. We also discussed the prospective use of mutant *ppsA*, *pks12*, and *pks13* genes as drug targets.

KEYWORDS

Mycobacterium tuberculosis, mutation, polyketide synthases, transmission, phylogenetic analysis

1 Introduction

Tuberculosis remains a major cause of suffering worldwide. Globally in 2020, tuberculosis was the second leading cause of death from infectious disease in humans worldwide, following COVID-19. Approximately 10 million individuals contracted tuberculosis disease, and roughly 1.5 million lost their lives. (Global tuberculosis report 2021, 2021). Successful TB transmission depends on the interplay of human behavior, host immune responses, and *Mycobacterium tuberculosis (Mtb)* virulence factors. More attention has been paid to the transmission of multidrug-resistant (MDR) or extensively drug-resistant (XDR) *Mtb*, (Clark et al., 2013; Yang et al., 2017a; Madikay et al., 2017; Bouzouita I Fau - Cabibbe et al., 2019; Dixit et al., 2019; Jiang et al., 2020a), or described the dynamics of TB transmission combined with host risk factors (Genestet C Fau - Tatai et al., 2019; Liu et al., 2021). To date, there has been no systematic study to delineate the role of virulence factors in TB transmission. (Global tuberculosis report 2021, 2021). In *Mtb*, polyketide synthase (PKS) genes encode large multifunctional proteins that contain all domains required to catalyze the various steps involved in the biosynthesis of complex mycobacterial lipids. These lipids have been shown to be key players for mycobacterial pathogenicity and transmissibility (Camacho et al., 1999; Cox et al., 1999; Asselineau et al., 2002; Reed et al., 2004; Tsenova et al., 2005; Astarie-Dequeker et al., 2009; Verschoor et al., 2012; Cambier et al., 2014; Passemar et al., 2014) and contributors to the cell envelope permeability barrier to antimicrobial drugs (Camacho et al., 2001; Alibaud et al., 2011; Chavadi et al., 2011; Yu et al., 2012).

Polyketide synthases are grouped into three protein structure-based types: Type I, Type II, and Type III. According to a previous study, Type I PKS generally synthesizes complex metabolites with the use of a modular or iterative biosynthetic mechanism (Gokhale et al., 2007a). In an iterative mechanism, the final product is produced by repeating the same active sites, while modular proteins follow an assembly-line mechanism (Gokhale et al., 2007a). This study primarily focused on three lipids: DIMs, MPMs, and mycolic acids and their corresponding synthesis proteins, ppsA, pks12, and pks13, respectively. PpsA, pks13 and pks12 were belong to Type I PKS. PpsA and pks13 belong to modular I PKS, while pks12 belongs to iterative I PKS (Onwueme et al., 2005). Dimycocerosates are a family of compounds that contain two diols, phthiocerol and phenolphthiocerol, which have been proven to be major mycobacterial virulence factors with complex molecular mechanisms of action (Camacho et al., 1999; Cox et al., 1999; Reed et al., 2004; Tsenova et al., 2005; Astarie-Dequeker et al., 2009; Cambier et al., 2014; Passemar et al., 2014). The clusters of *ppsABCDE* genes had been shown to be involved in the biosynthesis of phthiocerol products (Figure 1A). Phthiocerol products are synthesized by catalyzing a stepwise chain elongation and functional group modification with modular organization of pps proteins (Trivedi et al., 2005; Siméone et al., 2010). As shown in Figure 1C, the pks12 protein is involved in biosynthesis of a phospholipid MPM (Matsunaga et al., 2004). Recently, it has been discovered that novel phospholipid MPMs isolated from *Mtb* and other pathogenic mycobacteria consist of a mannosyl-β-1-phosphate. Mycolic acids are key players in the infectious process (Moody DB et al., 2002; Geisel RE et al., 2005; Layre et al., 2009; Esin et al., 2013). In mycolic acid synthesis, pks13 performs Claisen condensation of a C26 α-alkyl branch and C40–60 meromycolate precursors as the final assembly stage (Figure 1B) (Portevin et al., 2004). It has been demonstrated that this activity is crucial both *in vitro* and *in vivo* (Portevin et al., 2004; Wilson et al., 2013). Additionally, according to several genomic investigations, some PKS disruption mutants in mycobacteria have altered lipid profiles and some also show virulence attenuation (Sirakova et al., 2001; Dubey et al., 2002). PKS proteins play a significant role in enhancing the virulence and pathogenicity of *M.tb*. Nonetheless, the exact regulatory mechanism of PKS in *M.tb* is still unclear, and there is limited research on how gene mutations affecting PKS impact the transmission of *M.tb*. Thus, to develop effective TB control strategies, it is also necessary to gain a deeper understanding of the role of PKS gene in TB transmission. Therefore, this study aimed to identify the role of PKS mutations in TB transmission and assess its risk and characteristics. We also discussed the prospective use of mutant PKS genes as drug targets.

2 Materials and methods

2.1 Clinical isolates

Genomic DNA was successfully extracted from 1,468 *Mtb* samples from Shandong Provincial over a 5-year period for this study, and a total of 1,449 samples passed quality control (QC). Quality control of sequenced reads was carried out using FastQC software. In this study, we combined the 1,449 *Mtb* whole genome dataset with another genome dataset consisting of 1755 isolates, which were acquired from nine previously published articles (Zhang et al., 2013; Luo et al., 2015; Yang et al., 2017b; Liu et al., 2018a; Hicks et al., 2018; Yang et al., 2018; Chen et al., 2019; Huang et al., 2019; Jiang et al., 2020b). These samples were randomly collected from 21 provinces, 4 municipalities, and 5 autonomous regions in China, totaling to 3,204 isolates, from 2011 to 2019, to analyze the role of PKS mutation in TB transmission. Of the 3,204 *Mtb* isolates, Shandong contributed the most isolates (1,484), Yunnan the fewest (2), Xinjiang and Hainan (3), Qinghai and Tianjin (5), Gansu (8), Chongqing (9) and other provinces, municipalities, or autonomous regions contributed from 11 to 454 isolates; 73 had undetermined sources (Figure 2). We added a Supplementary Table S1) of the list of the1755 isolates, together with their corresponding meta-data. We also added a flowchart (Figure 3) about the process of identification and exclusion of genomic data.

2.2 Whole-genome sequencing and SNP identification

The genome was sequenced using HiSeq 4,000 (Illuminia Inc., San Diego, CA, United States). We discarded low-quality raw reads from paired-end sequencing. Maximal Exact Match algorithm was implemented by bwa mem (version 0.7.17-r1188) and was used to align the read to the H37Rv reference genome (NC_000962). Samclip (version 0.4.0) and samtools markdup (version 1.15) were used to remove clamped alignments and duplicated reads, excluding samples

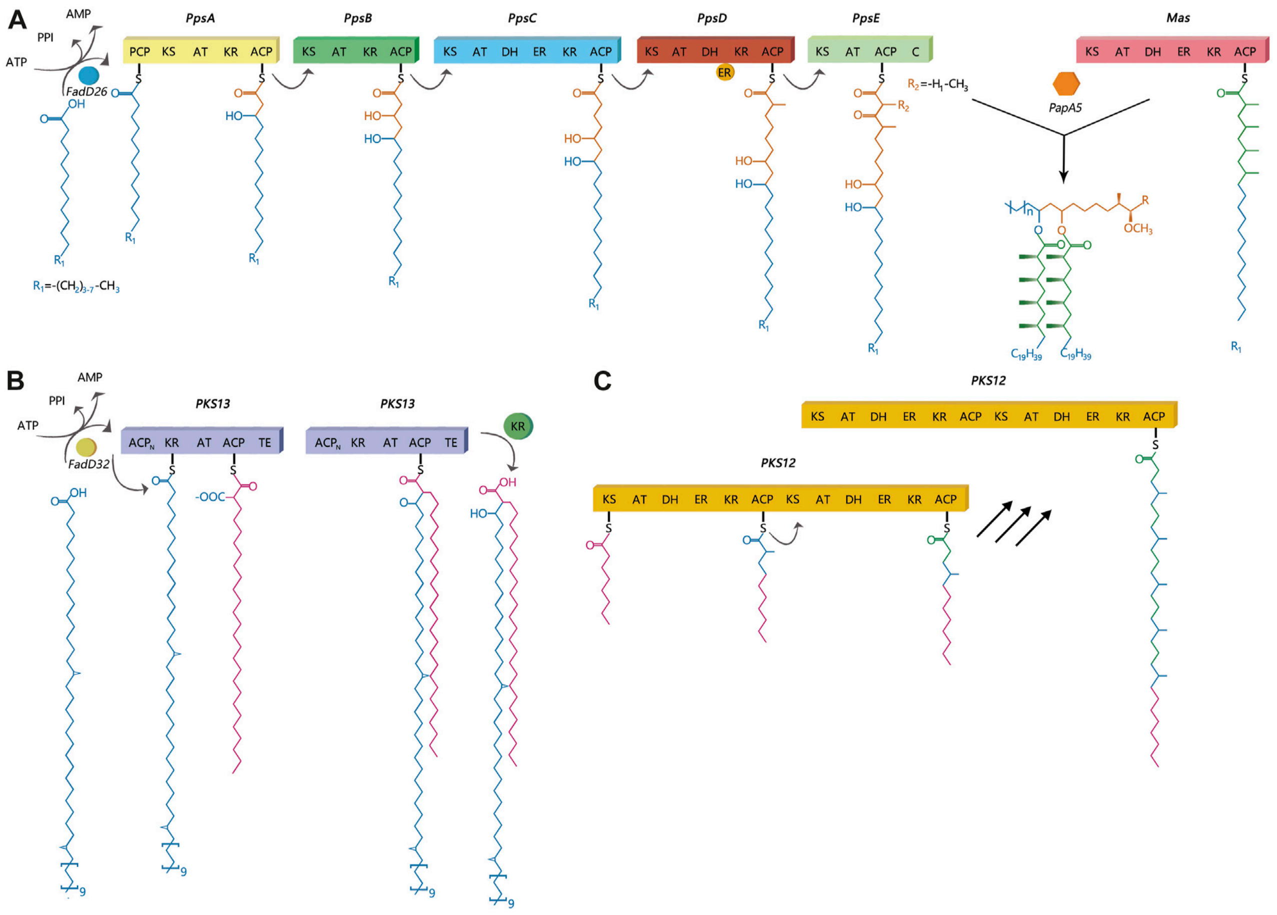


FIGURE 1
Catalytic and mechanistic versatility of ppsA, pks12, pks13. **(A)** PpsA initiates biosynthesis of phthiocerol products. It does this by extending its substrate using a malonyl-CoA extender unit; the same has been observed for ppsB and ppsC proteins. The ppsD and ppsE proteins add two (R)-methylmalonyl units to the substrate. **(B)** The pks13 protein consists of five domains, including two acyl carrier protein domains, a β-ketoacyl-synthase, an acyltransferase, and a C-terminal thioesterase (TE) domain, which together contain all the activities required for the condensation of two long-chain fatty acids. **(C)** There are two complete sets of modules in pks12, which produce mycoketide using five alternating condensations of methylmalonyl and malonyl units. The iterative process would generate a fully saturated chain with branching at each alternate ketide unit.

with coverage less than 98% and depth less than 20 (Li and Durbin, 2009; Li et al., 2009; Li and Durbin, 2010; Li, 2013). Variant calling was performed using Freebayes (version 1.3.2) and bcftools (version 1.15.1) with a filter parameter 'FMT/GT = "1/1"&& QUAL>=100 && FMT/DP>=10 && (FMT/AO)/(FMT/DP)>=0'. Single nucleotide polymorphisms in previously defined repetitive regions were excluded, including *PPE* and *PE-PGRS* genes, and mobile elements or repeat regions and repeat bases generated by TRF (version 4.09) and Repeatmask (version 4.1.2-p1) (Benson, 1999; Saha et al., 2008; Garrison and Marth, 2012; Danecek et al., 2021). The filtered vcf file was annotated using snpEff (version 4.3t) to get the final SNP samples (http://SnpEff.sourceforge.net/) (Cingolani et al., 2012). Genotypic drug resistance of each isolate was predicted in TBProfiler using an established library of mutations (https://github.com/jodyphelan/tbdb) (Coll et al., 2015). The virulence factor database (http://www.mgc.ac.cn/VFs/) contains various medically important bacterial pathogen virulence factors, which include 86 experimentally confirmed and 171 putative genes related to the virulence of *Mtb* (Liu et al., 2022). There are at least 24 different PKS encoded in the genome (Cole et al., 1998).

2.3 *Mtb* lineage and genomic cluster

We used the web-based tool TBProfiler (version 4.3.0) to analyze 3204 *M. tuberculosis* WGS data to assign lineages and predict drug resistance (Phelan JE et al., 2019). Genomic clusters were ascertained independently of the epidemiological data, and Genomic clusters were inferred based on how genetically similar two isolates were from each other. The upper thresholds of genomic relatedness or cluster is defined as 12 SNPs or alleles cut off or less and a recent transmission event is defined as 5 or less SNPs or alleles (Walker et al., 2013; Kohl TA et al., 2018). If two isolates exhibited a distance of more than 12 SNPs or alleles, they were called unique strains. In this study *M. tuberculosis* isolates with a genomic difference (s) ≤ 10 single nucleotide polymorphisms (SNPs) were defined as a genomic cluster (Yang et al., 2017a) for further analysis of transmission cluster to avoid missing cases and incorporating recent and old transmission events, which is similar to definitions used in previous genomic studies of *M. tuberculosis* transmission (Walker et al., 2013; Walker et al., 2014; Guerra-Assunção et al., 2015). As suggested by recent analysis of intra-patient variation, the estimate of 5 SNPs may be too low (Lieberman TD

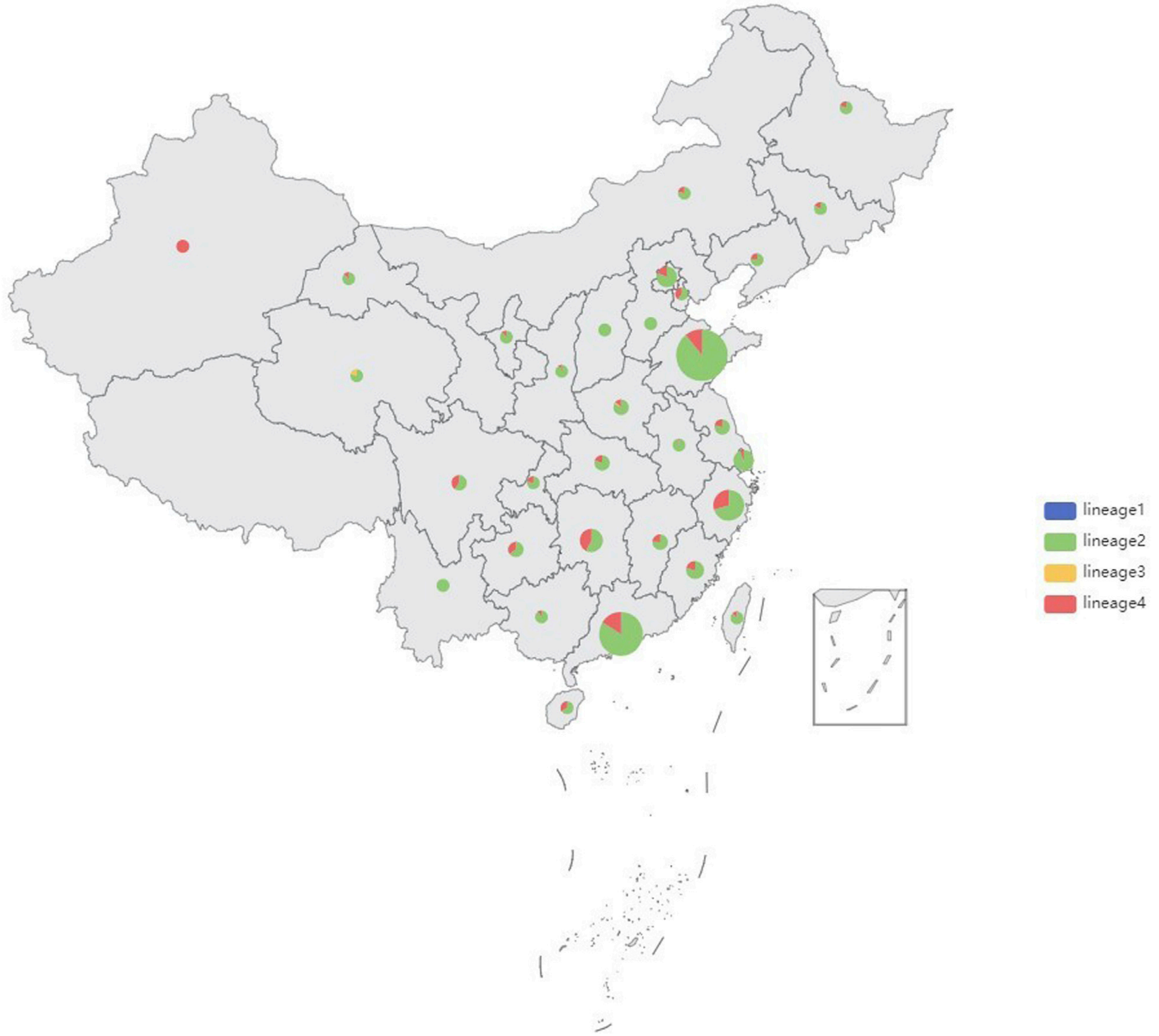


FIGURE 2
Sample size and lineages proportion in different regions of the 3,204 isolates, China, 2011–2019.

et al., 2016), we finally chose the cut-off of 10 SNPs to define transmission clusters for further analysis based on the previous study (Holt et al., 2018). The clustering was performed based on the statistical analysis which was not associated with sampling.

2.4 Phylogenetic analysis

Reference genome with only substitution variants instantiated was used as the sample's genome. Maximum-likelihood (ML) phylogenetic trees were constructed and dated by IQ-TREE (v1.6.12) model "JC + I + G4" with 1,000 ultrafast bootstrap replicates and treetime (v0.9.0) [GitHub - neherlab/treetime: Maximum likelihood inference of time stamped phylogenies and ancestral reconstruction. https://github.com/neherlab/treetime.] (Zelner et al., 2016)The trees were constructed using the highest likelihood model selected by automatic model selection in IQ-TREE (v1.6.12), which utilized the JC model of nucleotide substitution and invariable site plus discrete Gamma model of rate heterogeneity to analyze the genome samples with only substitution variants replaced in reference sequence. Sampling dates were used to construct a temporal phylogeny using TreeTime (v0.9.0) [GitHub - neherlab/treetime: Maximum likelihood inference of time stamped phylogenies and ancestral reconstruction. https://github.com/neherlab/treetime.] (Zelner et al., 2016), and tip-randomization was performed to confirm the presence of a strong temporal signal. Bayesian evolutionary analyses were conducted to identify the best substitution, clock, and demographic models, with marginal likelihood estimates used for model selection. The visualization of the bacteriological information was performed using Interactive Tree of Life (Version 6.6) (Letunic and Bork, 2021).

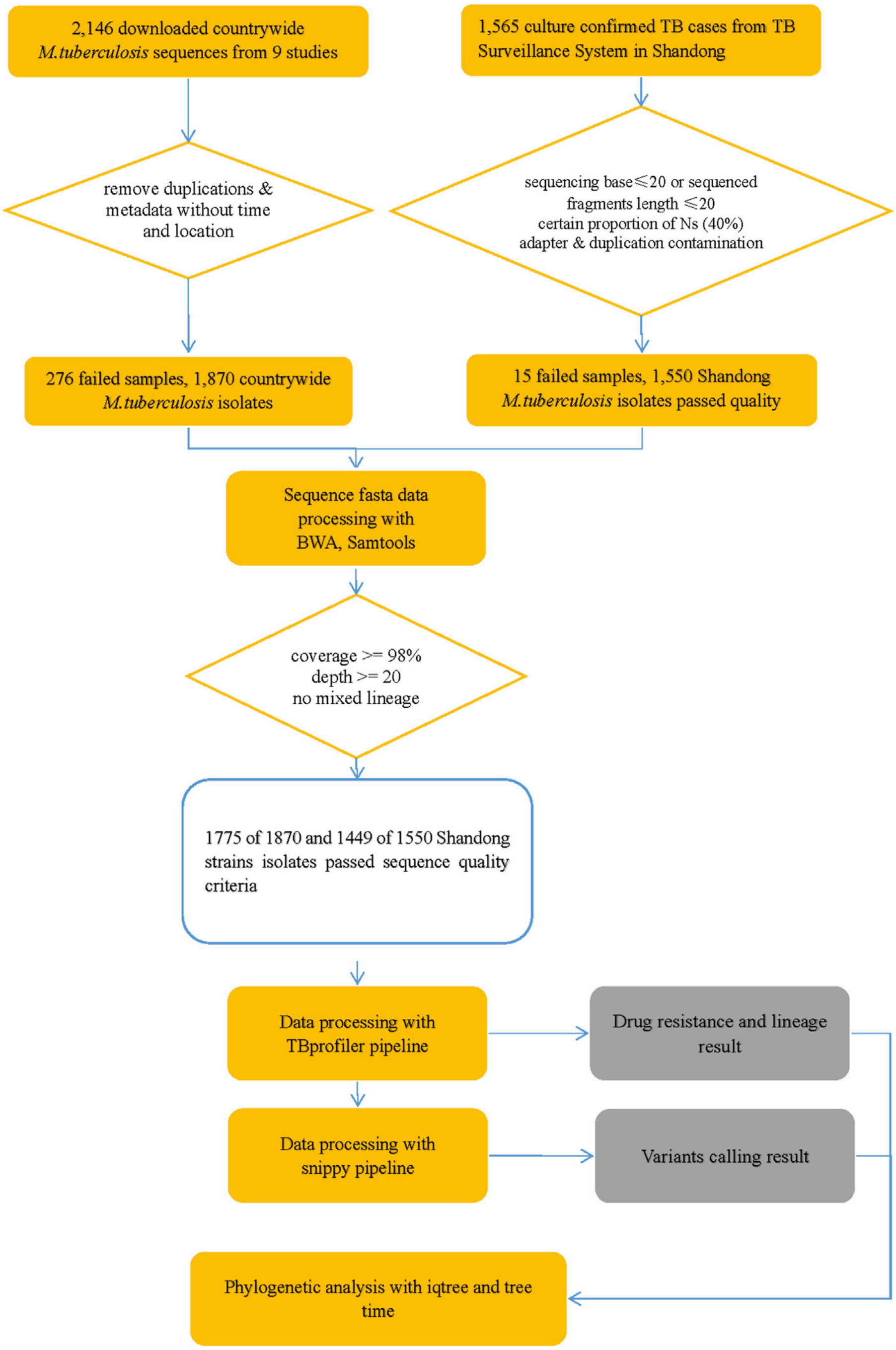

2,146 downloaded countrywide *M.tuberculosis* sequences from 9 studies
1,565 culture confirmed TB cases from TB Surveillance System in Shandong
remove duplications & metadata without time and location
sequencing base≤20 or sequenced fragments length ≤20 certain proportion of Ns (40%) adapter & duplication contamination
276 failed samples, 1,870 countrywide *M.tuberculosis* isolates
15 failed samples, 1,550 Shandong *M.tuberculosis* isolates passed quality
Sequence fasta data processing with BWA, Samtools
coverage >= 98% depth >= 20 no mixed lineage
1775 of 1870 and 1449 of 1550 Shandong strains isolates passed sequence quality criteria
Data processing with TBprofiler pipeline
Drug resistance and lineage result
Data processing with snippy pipeline
Variants calling result
Phylogenetic analysis with iqtree and tree time

FIGURE 3
Flowchart 1: a flowchart about the process of identification and exclusion of genomic data. *M tuberculosis*, *Mycobacterium tuberculosis*; TB, tuberculosis.

2.5 Statistical analysis

The mutation loci in the polyketide synthesis gene region between "clustered isolates" and "non-clustered isolates" was compared using univariate and multivariate logistic regression analysis in different lineages. Factors with a *p*-value less than 0.05 in the final model were considered to be independently associated with genomic clusters. The odds ratios (OR) and 95% confidence intervals (95% CI) were calculated. All statistical analyses were performed in R version 4.2.0 unless otherwise stated. Finally, a sensitivity analysis was performed to determine whether there was a rank correlation between cluster size and clustering rate with ordered logistic regression analysis. The R code see Supplementary Materials 2. Only fixed mutations (25%≤frequency<100%) were calculated from different lineages. The mutation frequency was calculated as the percentage of mutation isolates among the number of total isolates in different lineages. The detailed mutations were indicated in Table 3. The clustering rate was calculated as the percentage of cluster isolates among total isolates (number of cluster isolates/number of total

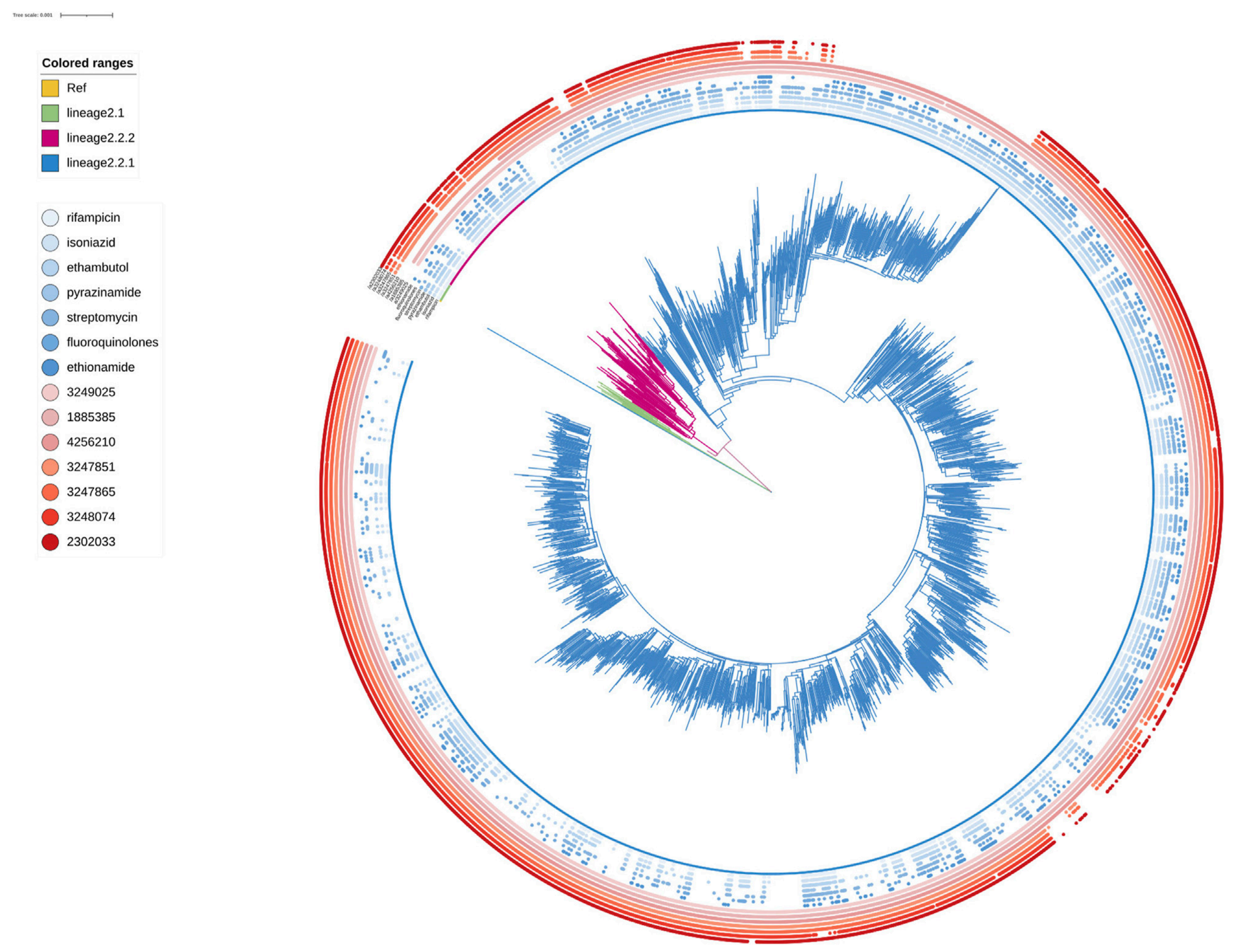


FIGURE 4
Phylogenetic tree for lineage2. Green, red and blue branches indicated L2.1, L2.2.2 and L2.2.1 strains, respectively. The inner blue dots indicated the resistance to known antimicrobial drugs. The outermost red dots showed the strains contained SMs.

isolates). Only nonsynonymous mutations were analyzed. Insertions and deletions were excluded from the analysis as they are often the result of errors in genome assembly. In terms of SNPs, isolates that possess the mutation in the PKS gene region are referred to as mutation isolates.

2.6 Predicted impact of mutations on proteins

Protein prediction algorithm, I-Mutant v2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html), was used to predict the functional impact of noteworthy SNPs on protein structure and function.

2.7 Genomic data availability

The newly sequenced whole genome dataset of 1,449 *M. tuberculosis* strains was deposited in the NCBI Bio Project (https://www.ncbi.nlm.nih.gov/sra/), and 1755 other isolates were downloaded from the European Nucleotide Archive repository (Supplementary Table S1). Additional data can be obtained by contacting the corresponding authors upon request.

3 Results

3.1 Genetic diversity

shown in the map (Figure 2), 85.73% (2,745/3,204) strains belonged to Lineage 2 (Beijing lineage), 13.84% (443/3,204) to Lineage 4 (Euro-American lineage), while only 0.31% (11/3,204) to Lineage 3 (East African-Indian lineage) and 0.12% (5/3,204) to Lineage 1(Indo-Oceanic lineage). A maximum likelihood phylogenetic tree was constructed for lineage 2 and lineage 4 *Mtb* isolates (Figure 4; Figure 5).

3.2 Clustering rate of the *Mtb* isolates

One thousand four hundred and sixty-four out of 2,745 isolates in lineage 2 were grouped into 446 genomic clusters (Table 1). The

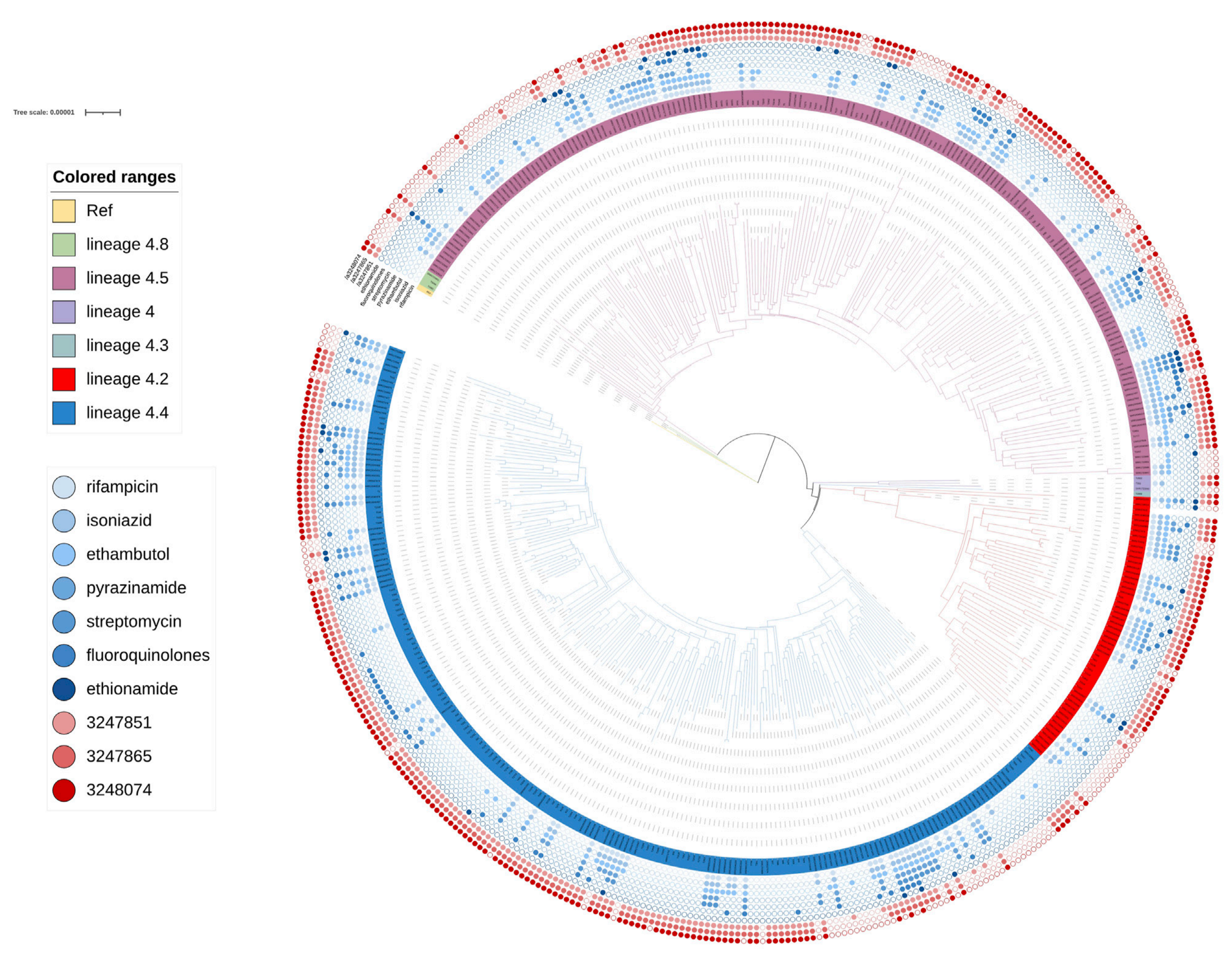


FIGURE 5
Phylogenetic tree for lineage4. Green, rose red, purple, dark green, red and blue branches indicated sublineage 4.8, sublineage 4.5, sublineage 4, sublineage 4.3 sublineage 4.2 and sublineage 4.4 strains, respectively. The inner blue dots indicated the resistance to known antimicrobial drugs. The outermost red dots showed the strains contained SMs.

clustering rate was 53.33%, which indicated the transmission of lineage 2 in China from 2011 to 2019. The genomic clusters consisted of 2–109 isolates. Majority of the clusters had two isolates, accounting for 36.47% (534/1,464). There were 52 genomic clusters consisting of two to nine isolates in lineage 4. The clustering rate of lineage 4 was 29.86%.

3.3 Drug resistance associated with genomic clusters

Known antimicrobial resistance mutations were detected in lineage 2 and lineage 4 (Table 2). Mutations in lineage 2 associated with resistance to rifampicin, isoniazid, pyrazinamide, streptomycin, ethambutol, fluoroquinolones, and ethionamide were all associated with genomic clusters ($p < 0.05$). This was the same as lineage 4, which was associated with resistance to streptomycin, isoniazid, rifampicin, and pyrazinamide and had a higher risk of clustering ($p < 0.05$). The phylogenetic trees show the drug resistance profile for 7 anti-TB drugs based on the presence of validated resistance-conferring mutations (Figure 4; Figure 5). Mutations occurred mainly in drug resistance genes such as *katG, rpoB, rpsL, embB, pncA, gyrA*, and *ethA*. Drug resistance is an important factor of TB transmission. In our study, we just used the Drug resistance mutations as exposure factors in multivariate logistic regression analysis to improve the sensitivity of analysis results.

3.4 Spread mutation (SM)

As shown in Table 3, the univariate logistic analysis detected eight loci mutations in the PKS gene region of L2 isolates, which were statistically significant ($p < 0.05$). Seven were risk factors (OR>1) and one was a protective factor (OR<1). The seven risk factors [ppsA(3,248,074, 3,247,851, 3,247,865, 3,249,025), pks12(2,302,033), pks13(4,256,210) and pks8(1,885,385)] were defined as Spread Mutations (SMs), meaning isolates with the seven SMs were more likely to be clustered than those without.

TABLE 1 The cluster size and the number of genomic clusters of the *Mycobacterium tuberculosis* isolates in lineage2 and lineage4.

No. of isolates in clusters	No. of clusters	No. of isolates	Proportion (%)
Lineage2			
0	0	1281	46.67
2	267	534	19.45
3 to 6	157	585	21.31
≥7	22	345	12.57
Total	446	2745	100
Lineage4			
0	0	310	70.14
2	35	70	15.84
3 to 6	16	53	11.99
≥7	1	9	2.04
Total	52	442	100

The basic information was shown in Table 4. All seven SMs were found in L2 and three SMs were found in L4 [ppsA(3248074,3247865,3,247,851)].

The clustering rate of lineage 2 was 53.33%, while lineage 4 was 29.86%. Lineage 2 exhibited a higher clustering rate than lineage 4 (Table 1), which was determined that the isolates of L2 spread faster than those of L4. The SNPs of lineage 2 and lineage 4 were not exactly the same. Some SNPs were found in lineage 2 but not found in lineage 4. The vast majority of these SNPs of lineage 2 exhibit high clustering rate (above 52.52%). Similarly, some SNPs were found in lineage 4 and not found in lineage 2. The clustering rate of these SNPs of lineage 4 ranged from 26.79% to 40.91%. We found seven SMs in lineage 2, while three SMs in lineage 4. However, owing to the smaller sample size of L4, we cannot guarantee that there were no hidden SMs. Interestingly, the clustering rate of SNP [pks12(2302033)] was higher than that of other SNP in lineage 4, but it was not statistically significant (p < 0.05) in univariate logistic analysis. We think it was because the amount of mutation isolates that contained SNP[pks12(2302033)] was too small.

We found four SMs in lineage 2 were statistically significant, while none in lineage 4 in multivariable regression analysis (Table 5). Due to the large standard error, P and OR were undetermined, and this can be due to the small sample size of lineage 4. In multivariable regression analysis, factors independently associated with genomic clusters including SMs and antimicrobial resistance mutations associated with genomic clusters of different lineages were introduced into the statistical model. *PpsA* (3249025), *pks12* (2302033) and *pks13* (4256210) are risk factors, while *ppsA* (3248074) was protect factor. Notably, the OR of *ppsA* (3249025) in lineage 2 were larger and the mutation was more likely to be clustered compared to other SMs.

Our study attempts to identify mutations that increase transmissibility. Lineage 2.2.1(Beijing lineage) strains are more transmissible than other *Mtb* lineages (Holt et al., 2018).

TABLE 2 Known antimicrobial resistance mutations associated with genomic clusters of lineage 2 and lineage 4.

Antimicrobial	No. of isolates	Clustering percetenge (%)	OR	P	Mutation genes
Lineage2					
rifampicin	1,203	43.83	1.897 (1.626, 2.211)	**<0.001**	*rpoB, rpoC*
isoniazid	1,264	46.05	1.897 (1.626, 2.211)	**<0.001**	*katG, fabG1, ahpC, inhA*
pyrazinamide	519	18.91	1.552 (1.276, 1.887)	**<0.001**	*pncA*
streptomycin	1,021	37.19	1.753 (1.497, 2.053)	**<0.001**	*rpsL, rrs, gid*
ethambutol	744	27.10	1.746 (1.500, 2.033)	**<0.001**	*embB, embA*
fluoroquinolones	458	16.68	1.565 (1.274, 1.924)	**<0.001**	*gyrA, gyrB*
ethionamide	363	13.22	1.267 (1.013, 1.585)	**0.038**	*fabG1, ethA, inhA*
Lineage4					
rifampicin	180	40.72	1.719 (1.139, 2.596)	**0.01**	*rpoB, rpoC*
isoniazid	180	40.72	1.719 (1.139, 2.596)	**0.01**	*katG, fabG1, ahpC*
pyrazinamide	109	24.66	1.786 (1.134, 2.814)	**0.012**	*pncA*
streptomycin	70	15.84	1.847 (1.091, 3.129)	**0.022**	*rpsL, rrs, gid*
ethambutol	189	42.76	1.333 (0.885, 2.009)	0.169	*embB, embA*
fluoroquinolones	75	16.97	1.507 (0.896, 2.534)	0.122	*gyrA, gyrB*
ethionamide	38	8.60	0.826 (0.389, 1.752)	0.618	*fabG1, ethA, inhA*

OR, odds ratio. The bold values mean these mutations were statistically significant.

TABLE 3 Univariate regression analysis on SMs associated with clustering in PKS gene region of lineage 2 and lineage 4[a]

	Lineage2					*Lineage4*				
Genomic position	No. of isolates	Mutation frequency	Clustering rate *	OR	P	No. of isolates	Mutation frequency	Clustering rate *	OR	P
ppsA										
3,248,074	2025	73.77%	54.56%	1.21 (1.02,1.43)	**0.03**	282	63.80%	33.33%	1.61 (1.04, 2.51)	**0.035**
3,247,851	2,298	83.71%	55.48%	1.70 (1.39,2.09)	**<0.001**	315	71.27%	33.02%	1.74 (1.09, 2.86)	**0.024**
3,247,865	2,279	83.02%	55.59%	1.71 (1.40,2.09)	**<0.001**	310	70.14%	33.55%	1.88 (1.17, 3.07)	**0.01**
3,249,025	2,723	99.19%	53.65%	7.33 (2.49,31.3)	**0.001**	0	0	0	0	0
3,247,316	2,733	99.56%	53.24%	0.38 (0.08,1.28)	0.15	440	99.55%	29.77%	0.42 (0.02, 10.80)	0.55
Pks12										
2,302,033	2,234	81.38%	55.73%	1.68 (1.38,2.04)	**<0.001**	22	5%	40.91%	1.67 (0.70,4.01)	0.25
2,296,042	2,730	99.45%	53.26%	0.57 (0.18,1.61)	0.31	250	56.56%	28.00%	0.82 (0.54, 1.23)	0.33
2,300,546	2,607	94.97%	52.89%	0.70 (0.49,0.99)	**0.047**	399	90.27%	28.82%	0.62 (0.33, 1.20)	0.15
2,296,297	0	0	0	0	0	112	25.34%	26.79%	0.82 (0.50, 1.31)	0.41
Pks13										
4,256,210	2,582	94.06%	54.11%	1.69 (1.23,2.34)	**0.001**	0	0	0	0	0
4,258,106	2,209	80.47	53.87%	1.12 (0.92,1.35)	0.25	0	0	0	0	0
Pks6										
485,810	2,738	99.75	53.25%	0.19 (0.01,1.11)	0.12	0	0	0	0	0
488,579	0	0	0	0	0	196	44.34%	29.59%	1.04 (0.69, 1.57)	0.84
Pks7										
1,877,744	2,744	99.96%	53.35%	#	0.95	0	0	0	0	0
1,881,343	0	0	0	0	0	185	41.86%	31.35%	1.13 (0.75, 1.70)	0.56
Pks8										
1,885,772	2,739	99.78%	53.27%	0.23 (0.01,1.42)	0.18	440	99.54%	30.00%	#	0.98
1,885,385	2,718	99.02%	53.57%	2.74 (1.24,6.66)	**0.017**	0	0	0	0	0
Pks15										
3,296,371	1803	65.68%	52.52%	0.91 (0.78,1.07)	0.24	0	0	0	0	0
3,296,843	2,728	99.38%	53.19%	0.35 (0.10,0.99)	0.066	440	99.55%	29.77%	0.42 (0.02, 10.80)	0.55

(Continued on following page)

TABLE 3 (*Continued*) Univariate regression analysis on SMs associated with clustering in PKS gene region of lineage 2 and lineage 4[a]

	Lineage2					*Lineage4*				
Genomic position	No. of isolates	Mutation frequency	Clustering rate *	OR	P	No. of isolates	Mutation frequency	Clustering rate *	OR	P
ppsD										
3,267,163	0	0	0	0	0	183	41.40%	31.69%	1.16 (0.77, 1.75)	0.48
Pks1										
3,295,663	0	0	0	0	0	127	28.73%	30.71%	1.06 (0.67, 1.65)	0.81
Pks3										
1,315,191	2,745	100%	53.33%	#	#	442	100%	29.86%	#	#
Pks5										
1,722,228	2,725	99.27%	53.36%	1.14 (0.47,2.80)	0.76	0	0	0	0	0

[a]sMs refer to seven loci mutations statistically significant ($p < 0.05$) which are risk factors in the PKS, gene region of lineage2 isolates. *Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962. * The clustering rate was calculated as the percentage of cluster isolates among total isolates (number of cluster isolates/number of total isolates). #means there is no result in statistical software or the result was too large and nonsense. OR, odds ratio. The bold values mean these mutations were statistically significant.

TABLE 4 The basic information of the SMs associated with clustering in PKS gene of lineage 2 and lineage 4.

Genomic position	Type	References	Variant	Gene
Lineage2				
3,248,074	mnp	GC	AT	ppsA
3,247,851	complex	GCCCGG	ACTCGC	ppsA
3,247,865	complex	GCAAA	TAGGG	ppsA
3,249,025	snp	T	G	ppsA
2,302,033	snp	G	A	pks12
4,256,210	snp	G	T	pks13
1,885,385	snp	T	G	pks8
Lineage4				
3,248,074	complex	GC	AT	ppsA
3,247,851	complex	GCCCGG	ACTCGC	ppsA
3,247,865	complex	GCAAA	TAGGG	ppsA

Genomic evidence for enhanced transmission of the Beijing lineage has been documented in Russia (associated with antimicrobial resistance) (Casali et al., 2014) and Malawi (independent of antimicrobial resistance) (Guerra-Assunção JA et al., 2015). We also analyzed the SMs in lineage 2.2.1strains (Table 6). There are four SMs in lineage 2.2.1 strains. Only one SM [*pks12* (2302033)] was statistically significant ($p < 0.05$) in multivariable regression analysis. And the data showed that the SMs in lineage 2.2.1 have higher clustering rate than other lineages which are predicted to be more transmissible.

Evolutionary convergence has previously been used as a signal of positive selection to identify mutations associated with antimicrobial resistance in Mtb (Hazbón et al., 2008; Farhat MR et al., 2013). We think it can also be used as a signal of positive selection to identify mutations associated with genomic clusters. We reasoned that SMs with high clustering rate contributing to the enhanced transmissibility of lineage 2 should also be result of positive selection that is detectable as convergent or parallel evolution. SMs showed an unexpectedly high level of convergence among lineage 2.2.1, suggesting the action of selection.

From the above, it can be concluded that *ppsA* (3,249,025), *pks12* (2,302,033) and *pks13* (4,256,210) of lineage 2 were the final and meaningful mutation sites screened in our study, based on the results and analysis.

3.5 Sensitivity analysis

In the sensitivity analysis, the lineage 2 and lineage 4 data were divided into four groups and then reanalyzed using an ordinal regression analysis. As shown in Table 1, the first, second, third, and fourth group included non-clustered isolates, small clusters containing two isolates, clusters containing 3 to 6 isolates, and clusters containing ≥7 isolates, respectively. Only the SMs that were statistically significant in the univariate analysis and were risk factors were included in the statistical model.

As show in Table 7, *ppsA* (3,249,025), *pks12* (2,302,033), and *pks13* (4,256,210) of L2 were statistically significant and were risk factors in the ordinal regression analysis. Interestingly, the results for the ordinal and multivariate regression analysis were the same. The P and OR results for lineage 4 were undetermined, this can be attributed to the large standard error. The SM *ppsA*(3,249,025) was also more likely to be clustered than other SMs. Compared with non-clustered and small isolates, the larger and largest clustered isolates had higher clustering rate in the *ppsA* (3,249,025), *pks12*

TABLE 5 Multivariable regression analysis on SMs associated with clustering in PKS gene region of lineage 2 and lineage 4.

Genomic position	P	Or (95%CI)
Lineage2		
3,249,025	**0.005**	37.743 (3.060, 465.584)
2,302,033	**<0.001**	2.251 (1.487, 3.408)
4,256,210	**0.006**	1.643 (1.154, 2.340)
3,247,865	0.642	1.229 (0.515, 2.934)
3,248,074	**0.042**	0.742 (0.557, 0.989)
3,247,851	0.907	0.951 (0.411, 2.201)
1,885,385	0.07	0.133 (0.015, 1.179)
rifampicin	**0.001**	1.749 (1.271, 2.407)
pyrazinamide	0.462	0.904 (0.691, 1.183)
streptomycin	0.074	1.234 (0.980, 1.554)
fluoroquinolones	0.144	1.202 (0.939, 1.539)
ethambutol	0.752	0.957 (0.729, 1.257)
isoniazid	0.514	1.109 (0.813, 1.514)
ethionamide	0.176	0.834 (0.640, 1.085)
Lineage4		
3,247,865	0.999	a
3,248,074	0.832	1.096 (0.471, 2.550)
3,247,851	0.999	a
rifampicin	0.313	1.342 (0.758, 2.377)
pyrazinamide	0.591	1.211 (0.602, 2.436)
streptomycin	0.506	1.272 (0.626, 2.586)

[a]Means there is no result in statistical software or the result was too large and nonsense. OR, odds ratio. The bold values mean these mutations were statistically significant.

(2,302,033), and *pks13* (4,256,210) genes. The sensitivity analysis results did not change significantly compared to those of the univariate and multivariate regression analysis. The results of ordinal regression analysis based on the size of clustered isolates were like the main findings: SMs [*ppsA* (3,249,025), *pks12* (2,302,033), and *pks13* (4,256,210)] were risk factors for TB transmission.

3.6 Deleterious effect of SMs on proteins

The SMs were predicted to negatively affect the respective proteins that affect the protein instability in nearby structural areas (Table 7). We also checked the Uniprot database for the protein domain where the mutation occurs according to the protein sequence (Trivedi et al., 2005; Siméone et al., 2010). P*psA* (3,249,025) and *pks13* (4,256,210) occurs in linker, while *pks12* (2,302,033) occurs in active site. Linker was found to be the noncatalytic protein domain that connects different functional proteins.

4 Discussion

Genetic diversity analysis revealed that the majority of these isolates belonged to lineage 2(the predominant sublineage was 2.2.1), with lineage 4 accounting for a significant proportion, while lineage 3 and lineage 1were less frequent. In addition, lineage 2 exhibited a higher clustering rate compared to lineage 4. These findings suggest that Beijing strains were more geographically dispersed compared to lineage 4, which are consistent with previous research (van Soolingen et al., 1995; Pang Y et al., 2012; Liu et al., 2018b). The overwhelming majority of TB cases in China were caused by L2 and L4 strains. The result of analysis also reminds us of the need to prioritize resources in cases where contact tracing is most likely to yield results. In China, it may be beneficial to direct contact tracing resources to lineage 2 and lineage 4 cases, as they pose the greatest risk of onward transmission resulting in new active TB cases.

We identified three SMs of lineage 2 in the *ppsA* (3,249,025), *pks12* (2,302,033), and *pks13* (4,256,210) gene regions that can potentially improve TB transmission. These SMs were predicted to alter the function of their respective proteins, supporting the hypothesis that they may affect TB transmission. Several biological and biochemical studies have determined the importance of the identified genes, which have proved critical to the virulence of *Mtb* in several animal studies (Kondo E Fau - Kanai and Kanai, 1972; Kolattukudy et al., 1997; Glickman and Jacobs, 2001; Sirakova et al., 2003). Furthermore, the results of this study are supported by previous genomic epidemiological articles (Onwueme et al., 2005; Trivedi et al., 2005; Gokhale et al., 2007b; Chopra et al., 2008; Quadri, 2014).

The *ppsA* gene is one of the clusters of *ppsABCDE* genes that has been shown to be involved in the biosynthesis of phthiocerol products (Figure 1A). The biosynthesis of phthiocerol products requires almost 24 catalytic activities on five large multifunctional modular proteins (Trivedi et al., 2005). Thus, if there is a mutation in one of the *pps* genes that can change protein function, it may increase or decrease the efficiency of this specificity of hand-to-hand transfer of the chain from one pps protein to another. The pks12 protein is involved in biosynthesis of a phospholipid MPM (Matsunaga et al., 2004). A study by Sirakova et al. (2003) showed that the growth and virulence of mutant pks12 was attenuated in an *in vivo* murine model (Sirakova et al., 2003). In mycolic acid synthesis, ps13 performs Claisen condensation of a C26 α-alkyl branch and C40–60 meromycolate precursors as the final assembly stage (Portevin et al., 2004). According to Alland

TABLE 6 Ordinal regression analysis on SMs associated with clustering in PKS gene region.

Genomic position*	Value	Std.Error	T value	Ordered analysis	
				Or (95% CI)	P
Lineage 2					
ppsA					
3,249,025	3.5	0.91	3.86	33.069 (5.914,220.171)	**<0.001**
3,248,074	−0.41	0.12	−3.39	0.665 (0.525,0.842)	<0.001
3,247,865	0.41	0.39	1.04	1.505 (0.703,3.329)	0.15
3,247,851	−0.08	0.38	−0.22	0.92 (0.422,1.929)	0.415
Pks12					
2,302,033	0.68	0.19	3.49	1.973 (1.351,2.901)	**<0.001**
Pks13					
4,256,210	0.33	0.17	1.97	1.389 (1.005,1.934)	**0.024**
Pks8					
1,885,385	−1.71	0.64	−2.68	0.181 (0.051,0.661)	**<0.001**
Lineage 4					
ppsA					
3,248,074	0.15	0.39	0.38	a	a
3,247,865	7.92	35.38	0.22	a	a
3,247,851	−7.48	35.38	−0.21	a	a

[a]The standard error of regression coefficient in Lineage4 was too large. The bold values mean these mutations were statistically significant.

TABLE 7 Deleterious effect of SMs on PKS proteins[a].

Genomic position*	Nucleotide change	Amino acid change	Protein prediction	Protein domain*
ppsA				
3,249,025	T=>G	L1194R	Large decrease of stability	Linker*
3,248,074	GC=>AT	R877H	Large decrease of stability	Acyltransferase
3,247,865	GCAAA=>TAGGG	AQN807ARD	Large decrease of stability	Acyltransferase
3,247,851	GCCCGG=>ACTCGC	AR803TR	Large decrease of stability	Acyltransferase
Pks8				
1,885,385	T=>G	L1228V	Large decrease of stability	Linker*
Pks12				
2,302,033	G=>A	R1652H	Large decrease of stability	Enoyl reductase 1
Pks13				
4,256,210	G=>T	A1646S	Large decrease of stability	Linker*

[a]Functional impact of the SMs on protein structure and function was predicted on one protein prediction algorithms, I-Mutant v2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html). *Linker is the noncatalytic protein domain that connects different functional proteins. *Protein domain where the mutation occurs was checked in Uniprot database according to the protein sequence.

et al. (2000), there is a novel class of thiophenes that prevent fatty acyl-AMP loading on pks13, interfere with mycolic acid biosynthesis, and have bactericidal effects on *Mtb* (Alland et al., 2000; Wilson et al., 2013). Aggarwal et al. (2017) found a novel benzofuran class lead molecule that targets *pks13* with fantastic drug-like characteristics and excellent pharmacokinetic and safety features that are active against MDR and XDR *Mtb* clinical strains *in vitro*.

In addition, we predicted the impact of SMs on protein structure. Mutations in *ppsA*, *pks12*, and *pks13* genes affect instability in nearby structural areas, which may affect nearby biological functions. Modular PKSs are multidomain proteins. Each module contains at least three essential domains, which are catalytic sites or active sites, namely, acyl transferase (AT), acyl carrier protein (ACP), and keto synthase (KS) domains. These catalytic sites or active sites are interconnected by small stretches of relatively unconserved sequences called linkers, which are more than covalent connectors (Gokhale and Khosla, 2000). Some SMs occur on active sites while others occur on linkers. Apparently, if the mutation occurs at active sites, it can affect the function of the *pks* gene. New progress has shown that linkers play a strong role in building the structural and functional assemblies of these diverse modular proteins in signal transduction and polyketide biosynthesis (Briggs and Smithgall, 1999; Gokhale et al., 1999; Xu et al., 1999; Gokhale and Khosla, 2000). Chopra et al. (2008) found that these linkers play an important role in the formation of docking domains through interacting helices. This study also showed that single amino acid substitutions in the linkers had an effect on the catalytic rates of product formation (Chopra et al., 2008). Similar studies based on the erythromycin PKS have shown the crucial role of single amino acids in forming a docking complex (Weissman, 2006). Thus, if the mutation occurs in linkers, it can also have an impact on protein-protein interactions and affect catalysis (Chopra et al., 2008). Since the positions of the modules can be changed by suitable linker engineering (Gokhale et al., 1999), it is worth studying the mechanism of linker action in chemical biology.

In conclusion, this study presents evidence through statistical analysis that three *Mtb PKS* genes in lineage 2 may contribute to disease progression and higher transmission of certain strains. Previous studies suggest that virulence change is caused not by mass nonsynonymous mutations, but rather by several critical mutations that affect gene product activity (Hershberg et al., 2008; Mikheecheva et al., 2017). Distinct lipids in the cell wall of mycobacteria synthesized by the three genes are critical to the pathogen's ability to survive in the host's hostile environment. Their production involves a complex process that requires many enzymes (Mehra et al., 1984; Chan et al., 1989; Vachula et al., 1989). When these lipids are lost due to mutation, *M. tuberculosis* becomes less virulent in the host (Camacho et al., 1999; Cox et al., 1999). This process offers multiple ways to intervene in lipids production and thus opens up many possibilities for designing antimycobacterial agents. It might be possible to view the three SMs as specific targets for the development of medications for the treatment of mycobacteria-related infections in people. Notably, the OR of *ppsA* (3,249,025) in lineage 2 were larger and the mutation was more likely to be clustered compared to other SMs. Perhaps we should pay more attention to SNP: ppsA (3,249,025) in the following study. The SNP [ppsA (3,249,025)] should be further evaluated with animal and immunological experiments to test its importance regarding biological impact and as a new drug target.

5 Strength and limitations

This study has several limitations. First, we did not conduct animal and immunological experiments to find biological support for the SMs identified in this study. Second, we lack key host factors that may influence disease transmissibility, such as age, host immune status, and pulmonary cavitation, to rule out the effect of confounding factors, which could reveal independent effects of SMs influencing transmissibility. Finally, for the small sample size of lineage 4, hidden mutation sites may not be screened out. We cannot tell if the SMs of lineage 4 and lineage 2 were the same or different. Of course, the sample size of lineage 2 is large enough. The SMs we found were more reliable, which could provide credible data for TB prevention and treatment.

Data availability statement

The newly sequenced whole genome dataset of 1,449 *M. tuberculosis* strains has been submitted to the NCBI (https://www.ncbi.nlm.nih.gov/) under the accession number PRJNA1002108. 1755 other isolates were acquired from nine previously published articles (Supplementary Table S1). Additional data can be obtained upon request by contacting the corresponding authors.

Ethics statement

This study complies with the Declaration of Helsinki, and was approved by the Ethics Committee of Shandong Provincial Hospital, affiliated with Shandong University (SPH) and the Ethics Committee of Shandong Provincial Chest Hospital (SPCH), which waived informed patient consent because all patient records and information were anonymized and deidentified before the analysis.

Author contributions

HC-L, T-TW, YH, Y-FL, and YL conceived and designed the study. HC-L, T-TW, X-LK, and YH directed its implementation including the data analysis and writing of the paper. T-TW, and YH analyzed the data; YL, X-LK, Y-ML, Y-YL, PS, D-XW, L-LL, Y-ZZ, Q-LH., XZ, and Q-QA contributed materials/analytic tools; T-TW, Y-FL, and HC-L wrote and revised the manuscript. All authors reviewed and approved the manuscript.

Funding

This work was supported by the Department of Science & Technology of Shandong Province (CN) (Nos. 2007GG30002033 and 2017GSF218052), Natural Science Foundation of Shandong Province (CN) (No. ZR2020KH013 and ZR2021MH006), and Jinan Science and Technology Bureau (CN) (No. 201704100).

Acknowledgments

We sincerely appreciate all those who have generously offered to participate in our studies. We would like to express our gratitude to the hard-working research team who collected valuable field data. Their dedication and cooperation have been essential to the success of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1217255/full#supplementary-material

References

Alibaud, L., Rombouts, Y., Trivelli, X., Burguière, A., Cirillo, S. L. G., Cirillo, J. D., et al. (2011). A Mycobacterium marinum TesA mutant defective for major cell wall-associated lipids is highly attenuated in *Dictyostelium discoideum* and zebrafish embryos. *Mol. Microbiol.* 80 (1365-2958), 919–934. (Electronic)). doi:10.1111/j.1365-2958.2011.07618.x

Alland, D., Steyn, A. J., Weisbrod, T., Aldrich, K., and Jacobs, W. R. (2000). Characterization of the *Mycobacterium tuberculosis* iniBAC promoter, a promoter that responds to cell wall biosynthesis inhibition. *J. Bacteriol.* 182 (7), 1802–1811. doi:10.1128/jb.182.7.1802-1811.2000

Asselineau, C., Asselineau, J., Lanéelle, G., and Lanéelle, M. A. (2002). The biosynthesis of mycolic acids by Mycobacteria: current and alternative hypotheses. *Prog. Lipid Res.* 41 (0163-7827), 501–523. (Print)). doi:10.1016/s0163-7827(02)00008-5

Astarie-Dequeker, C., Le Guyader, L., Malaga, W., Seaphanh, F. K., Chalut, C., Lopez, A., et al. (2009). Phthiocerol dimycocerosates of *M. tuberculosis* participate in macrophage invasion by inducing changes in the organization of plasma membrane lipids. *PLoS Pathog.* 5, e1000289. 1553-7374 (Electronic)). doi:10.1371/journal.ppat.1000289

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27 (0305-1048), 573–580. (Print)). doi:10.1093/nar/27.2.573

Bouzouita, I., Cabibbe, A. M., Trovato, A., Daroui, H., Ghariani, A., Midouni, B., et al. (2019). Whole-Genome sequencing of drug-resistant *Mycobacterium tuberculosis* strains, tunisia, 2012–2016. *Emerg. Infect. Dis.* 25 (3), 538–546. doi:10.3201/eid2503.181370

Briggs, S. D., and Smithgall, T. E. (1999). SH2-kinase linker mutations release Hck tyrosine kinase and transforming activities in Rat-2 fibroblasts. *J. Biol. Chem.* 274 (0021-9258), 26579–26583. (Print)). doi:10.1074/jbc.274.37.26579

Camacho, L. R., Constant, P., Raynaud, C., Laneelle, M. A., Triccas, J. A., Gicquel, B., et al. (2001). Analysis of the phthiocerol dimycocerosate locus of *Mycobacterium tuberculosis*. Evidence that this lipid is involved in the cell wall permeability barrier. *J. Biol. Chem.* 276 (0021-9258), 19845–19854. (Print)). doi:10.1074/jbc.M100662200

Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B., and Guilhot, C. (1999). Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* 34 (0950-382X), 257–267. (Print)). doi:10.1046/j.1365-2958.1999.01593.x

Cambier, C. J., Takaki, K. K., Larson, R. P., Hernandez, R. E., Tobin, D. M., Urdahl, K. B., et al. (2014). Mycobacteria manipulate macrophage recruitment through coordinated use of membrane lipids. *Nature* 505 (1476-4687), 218–222. (Electronic)). doi:10.1038/nature12799

Casali, N., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., et al. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* 46, 279–286. doi:10.1038/ng.2878

Chan, J., Fujiwara, T., Brennan, P., McNeil, M., Turco, S. J., Sibille, J. C., et al. (1989). Microbial glycolipids: possible virulence factors that scavenge oxygen radicals. *Proc. Natl. Acad. Sci. U. S. A.* 86 (0027-8424), 2453–2457. (Print)). doi:10.1073/pnas.86.7.2453

Chavadi, S. S., Edupuganti, U. R., Vergnolle, O., Fatima, I., Singh, S. M., Soll, C. E., et al. (2011). Inactivation of tesA reduces cell wall lipid production and increases drug susceptibility in mycobacteria. *J. Biol. Chem.* 286 (1083-351X), 24616–24625. (Electronic)). doi:10.1074/jbc.M111.247601

Chen, X., Wang, S., Lin, S., Chen, J., and Zhang, W. (2019). Evaluation of whole-genome sequence method to diagnose resistance of 13 anti-tuberculosis drugs and characterize resistance genes in clinical multi-drug resistance *Mycobacterium tuberculosis* isolates from China. *Front. Microbiol.* 10, 1741. doi:10.3389/fmicb.2019.01741

Chopra, T., Banerjee, S., Gupta, S., Yadav, G., Anand, S., Surolia, A., et al. (2008). Novel intermolecular iterative mechanism for biosynthesis of mycoketide catalyzed by a bimodular polyketide synthase. *PLoS Biol.* 6, e163. 1545-7885 (Electronic)). doi:10.1371/journal.pbio.0060163

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly. (Austin)* 6, 80–92. 1933-6942 (Electronic)). doi:10.4161/fly.19695

Clark, T. G., Mallard, K., Coll, F., Preston, M., Assefa, S., Harris, D., et al. (2013). Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *Plos One* 8 (12), e83012. doi:10.1371/journal.pone.0083012

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544. 0028-0836 (Print)). doi:10.1038/31159

Coll, F., McNerney, R., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., et al. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 7, 51. 1756-994X (Print)). doi:10.1186/s13073-015-0164-0

Cox, J. S., Chen, B., McNeil, M., and Jacobs, W. R. (1999). Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* 402, 79–83. 0028-0836 (Print). doi:10.1038/47042

Danecek, P., Bonfield, P., Danecek, J. K., Liddle, J., Marshall, J., Ohan, V., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10 (2), giab008. doi:10.1093/gigascience/giab008

Dixit, A.A.-O., Freschi, L., Vargas, R., Calderon, R., Sacchettini, J., Drobniewski, F., et al. (2019). Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. *Sci. Rep.* 9, 5602. 2045-2322 (Electronic). doi:10.1038/s41598-019-41967-8

Dubey, V. S., Sirakova, T. D., and Kolattukudy, P. E. (2002). Disruption of msl3 abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in *Mycobacterium tuberculosis* H37Rv and causes cell aggregation. *Mol. Microbiol.* 45 (5), 1451–1459. doi:10.1046/j.1365-2958.2002.03119.x

Esin, S., Counoupas, C., Aulicino, A., Brancatisano, F. L., Maisetta, G., Bottai, D., Di Luca, M., et al. (2013). Interaction of *Mycobacterium tuberculosis* cell wall components with the human natural killer cell receptors NKp44 and Toll-like receptor 2. *Scand J Immunol* 77 (6), 460–9. doi:10.1111/sji.12052

Farhat Mr, S. B., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 45 (10), 1183–1189. doi:10.1038/ng.2747

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing[J]. *Quant. Biol.* doi:10.48550/arXiv.1207.3907

Geisel Re, S. K., Russell, D. G., and Rhoades, E. R. (2005). *In vivo* activity of released cell wall lipids of Mycobacterium bovis bacillus Calmette-Guérin is due principally to trehalose mycolates. *J. Immunol.* 174, 5007–5015. doi:10.4049/jimmunol.174.8.5007

Genestet, C., Tatai, C., Berland, J. L., Claude, J. B., Westeel, E., Hodille, E., et al. (2019). Prospective Whole-Genome Sequencing in Tuberculosis Outbreak Investigation, France, 2017-2018. *Emerg. Infect. Dis.* 25 (3), 589–592. doi:10.3201/eid2503.181124

Global tuberculosis report 2021 (2021). Available at: https://www.who.int/publications-detail-redirect/9789240037021.

Glickman, M. S., and Jacobs, W. R., Jr. (2001). Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline. *Cell.* 104 (0092-8674), 477–485. (Print)). doi:10.1016/s0092-8674(01)00236-7

Gokhale, R. S., and Khosla, C. (2000). Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol.* 4 (1367-5931), 22–27. (Print)). doi:10.1016/s1367-5931(99)00046-0

Gokhale, R. S., Sankaranarayanan R Fau - Mohanty, D., and Mohanty, D. (2007a). Versatility of polyketide synthases in generating metabolic diversity. *Curr. Opin. Struct. Biol.* 17 (0959-440X), 736–743. (Print)). doi:10.1016/j.sbi.2007.08.021

Gokhale, R. S., Saxena, P., Chopra, T., and Mohanty, D. (2007b). Versatile polyketide enzymatic machinery for the biosynthesis of complex mycobacterial lipids. *Nat. Prod. Rep.* 24 (0265-0568), 267–277. (Print)). doi:10.1039/b616817p

Gokhale, R. S., Tsuji, S. Y., Cane, D. E., and Khosla, C. (1999). Dissecting and exploiting intermodular communication in polyketide synthases. *Science* 284 (0036-8075), 482–485. (Print)). doi:10.1126/science.284.5413.482

Guerra-Assunção, J. A., Crampin, A. C., Houben, R. M., Mzembe, T., Mallard, K., Coll, F., et al. (2015). Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* 4, e05166. doi:10.7554/eLife.05166

Guerra-Assunção Ja, C. A., Houben, R. M., Mzembe, T., Mallard, K., Coll, F., Khan, P., et al. (2015). *Large-scale whole genome sequencing of* M. tuberculosis *provides insights into transmission in a high prevalence area*. Elife.

Hazbón, M. H., Motiwala, A. S., Cavatore, M., Brimacombe, M., Whittam, T. S., Alland, D., et al. (2008). Convergent evolutionary analysis identifies significant mutations in drug resistance targets of *mycobacterium tuberculosis*. *Antimicrob. Agents Chemother* 52 (9), 3369–3376. doi:10.1128/aac.00309-08

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., et al. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* 6 (1545-7885), e311. (Electronic)). doi:10.1371/journal.pbio.0060311

Hicks, N. D., Yang, J., Zhang, X., Zhao, B., Grad, Y. H., Liu, L., et al. (2018). Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance. *Nat. Microbiol.* 3 (9), 1032–1042. doi:10.1038/s41564-018-0218-3

Holt, K. E., McAdam, P., Thai, P. V. K., Thuong, N. T. T., Ha, D. T. M., Lan, N. N., et al. (2018). Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* 50 (6), 849–856. doi:10.1038/s41588-018-0117-9

Huang, H., Ding, N., Yang, T., Li, C., Jia, X., Wang, G., et al. (2019). Cross-sectional Whole-genome Sequencing and Epidemiological Study of Multidrug-resistant *Mycobacterium tuberculosis* in China. *Clin. Infect. Dis.* 69 (3), 405–413. doi:10.1093/cid/ciy883

Jiang, Q., Liu, Q., Ji, L., Li, J., Zeng, Y., Meng, L., et al. (2020a). Citywide transmission of multidrug-resistant tuberculosis under China's rapid urbanization: a retrospective population-based genomic spatial epidemiological study. *Clin. Infect. Dis.* 71 (1537-6591), 142–151. (Electronic)). doi:10.1093/cid/ciz790

Jiang, Q., Liu, Q., Ji, L., Li, J., Zeng, Y., Meng, L., et al. (2020b). Citywide transmission of multidrug-resistant tuberculosis under China's rapid urbanization: a retrospective population-based genomic spatial epidemiological study. *Clin. Infect. Dis.* 71 (1), 142–151. doi:10.1093/cid/ciz790

Kohl, T. A., Harmsen, D., Rothgänger, J., Walker, T., Diel, R., and Niemann, S. (2018). Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine* 34, 131–138. doi:10.1016/j.ebiom.2018.07.030

Kolattukudy, P. E., Fernandes, N. D., Azad, A. K., Fitzmaurice, A. M., and Sirakova, T. D. (1997). Biochemistry and molecular genetics of cell-wall lipid biosynthesis in mycobacteria. *Mol. Microbiol.* 24 (0950-382X), 263–270. (Print)). doi:10.1046/j.1365-2958.1997.3361705.x

Kondo E Fau - Kanai, K., and Kanai, K. (1972). Further demonstration of bacterial lipids in Mycobacterium bovis harvested from infected mouse lungs. *Jpn. J. Med. Sci. Biol.* 25 (0021-5112), 105–122. (Print)). doi:10.7883/yoken1952.25.105

Layre, E., Bastian, M., Mariotti, S., Czaplicki, J., Prandi, J., Mori, L., et al. (2009). Mycolic acids constitute a scaffold for mycobacterial lipid antigens stimulating CD1-restricted T cells. *Chem. Biol.* 16 (1), 82–92. doi:10.1016/j.chembiol.2008.11.008

Letunic, I.A.-O., and Bork, P. (2021). Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. 1362-4962 (Electronic). doi:10.1093/nar/gkab301

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv e-prints.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (1367-4811), 1754–1760. (Electronic)). doi:10.1093/bioinformatics/btp324

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26 (1367-4811), 589–595. (Electronic)). doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (1367-4811), 2078–2079. (Electronic)). doi:10.1093/bioinformatics/btp352

Lieberman Td, W. D., Misra, R., Xiong, L. L., Moodley, P., Cohen, T., Kishony, R., et al. (2016). Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nat. Med.* 22 (12), 1470–1474. doi:10.1038/nm.4205

Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50 (1362-4962), D912–D917. (Electronic)). doi:10.1093/nar/gkab1107

Liu, Q., Ma, A., Wei, L., Pang, Y., Wu, B., Luo, T., et al. (2018a). China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nat. Ecol. Evol.* 2 (12), 1982–1992. doi:10.1038/s41559-018-0680-6

Liu, Q., Ma, A., Wei, L., Pang, Y., Wu, B., Luo, T., et al. (2018b). China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nat. Ecol. Evol.* 2 (12), 1982–1992. doi:10.1038/s41559-018-0680-6

Liu, Q.A.-O., Liu, H., Shi, L., Gan, M., Zhao, X., Lyu, L. D., et al. (2021). Local adaptation of *Mycobacterium tuberculosis* on the Tibetan plateau. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2017831118–e2017836490. (Electronic)). doi:10.1073/pnas.2017831118

Luo, T., Comas, I., Luo, D., Lu, B., Wu, J., Wei, L., et al. (2015). Southern East asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with han Chinese. *Proc. Natl. Acad. Sci. U. S. A.* 112 (26), 8136–8141. doi:10.1073/pnas.1424063112

Madikay, S., Otu, J., Witney, A., Gehre, F., Doughty, E. L., Kay, G. L., et al. (2017). Whole-genome sequencing illuminates the evolution and spread of multidrug-resistant tuberculosis in Southwest Nigeria. *Plos One* 12 (9), e0184510. doi:10.1371/journal.pone.0184510

Matsunaga, I., Bhatt, A., Young, D. C., Cheng, T. Y., Eyles, S. J., Besra, G. S., et al. (2004). *Mycobacterium tuberculosis* pks12 produces a novel polyketide presented by CD1c to T cells. *J. Exp. Med.* 200, 1559–1569. 0022-1007 (Print)). doi:10.1084/jem.20041429

Mehra, V. F., Brennan, P. J., Rada, E., Convit, J., and Bloom, B. R. (1984). Lymphocyte suppression in leprosy induced by unique *M. leprae* glycolipid. *Nature* 308 (0028-0836), 194–196. (Print)). doi:10.1038/308194a0

Mikheecheva, N. E., Zaychikova, M. V., Melerzanov, A. V., and Danilenko, V. N. (2017). A nonsynonymous SNP catalog of *Mycobacterium tuberculosis* virulence genes and its use for detecting new potentially virulent sublineages. *Genome Biol. Evol.* 9 (1759-6653), 887–899. (Electronic)). doi:10.1093/gbe/evx053

Moody Db, B. V., Cheng, T. Y., Roura-Mir, C., Guy, M. R., Geho, D. H., Tykocinski, M. L., et al. (2002). Lipid length controls antigen entry into endosomal and nonendosomal pathways for CD1b presentation. *Nat. Immunol.* 3 (5), 435–442. doi:10.1038/ni780

Onwueme, K. C., Vos, C. J., Zurita, J., Ferreras, J. A., and Quadri, L. E. N. (2005). The dimycocerosate ester polyketide virulence factors of mycobacteria. *Prog. Lipid Res.* 44 (0163-7827), 259–302. (Print)). doi:10.1016/j.plipres.2005.07.001

Pang Y, Z. Y., Zhao, B., Liu, G., Jiang, G., Xia, H., Song, Y., et al. (2012). Spoligotyping and drug resistance analysis of *Mycobacterium tuberculosis* strains from national survey in China. *PLoS One* 7 (3), e32976. doi:10.1371/journal.pone.0032976

Passemar, C., Arbués, A., Malaga, W., Mercier, I., Moreau, F., Lepourry, L., et al. (2014). Multiple deletions in the polyketide synthase gene repertoire of *Mycobacterium tuberculosis* reveal functional overlap of cell envelope lipids in host-pathogen interactions. *Cell. Microbiol.* 16 (1462-5822), 195–213. (Electronic)). doi:10.1111/cmi.12214

Phelan Je, O. S. D., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S., O'Grady, J., et al. (2019). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 11 (1), 41. doi:10.1186/s13073-019-0650-x

Portevin, D., De Sousa-D'Auria, C., Houssin, C., Grimaldi, C., Chami, M., Daffé, M., et al. (2004). A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc. Natl. Acad. Sci. U. S. A.* 101 (0027-8424), 314–319. (Print)). doi:10.1073/pnas.0305439101

Quadri, L. E. (2014). Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. *Crit. Rev. Biochem. Mol. Biol.* 49 (1549-7798), 179–211. (Electronic)). doi:10.3109/10409238.2014.896859

Reed, M. B., Domenech, P., Manca, C., Su, H., Barczak, A. K., Kreiswirth, B. N., et al. (2004). A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 431 (1476-4687), 84–87. (Electronic)). doi:10.1038/nature02837

Saha, S., Bridges, S., Magbanua, Z. V., and Peterson, D. G. (2008). Empirical comparison of *ab initio* repeat finding programs. *Nucleic Acids Res.* 36 (1362-4962), 2284–2294. (Electronic)). doi:10.1093/nar/gkn064

Siméone, R., Léger, M., Constant, P., Malaga, W., Marrakchi, H., Daffé, M., et al. (2010). Delineation of the roles of FadD22, FadD26 and FadD29 in the biosynthesis of phthiocerol dimycocerosates and related compounds in *Mycobacterium tuberculosis*. *FEBS J.* 277 (1742-4658), 2715–2725. (Electronic)). doi:10.1111/j.1742-464X.2010.07688.x

Sirakova, T. D., Dubey, V. S., Kim, H. J., Cynamon, M. H., and Kolattukudy, P. E. (2003). The largest open reading frame (pks12) in the *Mycobacterium tuberculosis* genome is involved in pathogenesis and dimycocerosyl phthiocerol synthesis. *Infect. Immun.* 71 (0019-9567), 3794–3801. (Print)). doi:10.1128/iai.71.7.3794-3801.2003

Sirakova, T. D., Thirumala, A. K., Dubey, V. S., Sprecher, H., and Kolattukudy, P. E. (2001). The *Mycobacterium tuberculosis* pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis. *J. Biol. Chem.* 276 (20), 16833–16839. doi:10.1074/jbc.M011468200

Trivedi, O. A., Arora, P., Vats, A., Ansari, M. Z., Tickoo, R., Sridharan, V., et al. (2005). Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. *Mol. Cell.* 17, 631–643. 1097-2765 (Print)). doi:10.1016/j.molcel.2005.02.009

Tsenova, L., Ellison, E., Harbacheuski, R., Moreira, A. L., Kurepina, N., Reed, M. B., et al. (2005). Virulence of selected *Mycobacterium tuberculosis* clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. *Infect. Dis.* 192, 98–106. 0022-1899 (Print). doi:10.1086/430614

Vachula, M., Holzer Tj Fau - Andersen, B. R., and Andersen, B. R. (1989). Suppression of monocyte oxidative response by phenolic glycolipid I of *Mycobacterium leprae*. *J. Immunol.* 142 (0022-1767), 1696–1701. (Print)). doi:10.4049/jimmunol.142.5.1696

van Soolingen, D., de Haas, P. E., Douglas, J. T., Traore, H., Portaels, F., Qing, H. Z., et al. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. *Predominance a single genotype Mycobacterium Tuberc. Ctries. east Asia. J Clin Microbiol* 33 (12), 3234–3238. doi:10.1128/JCM.33.12.3234-3238.1995

Verschoor, J. A., Baird Ms Fau - Grooten, J., and Grooten, J. (2012). Towards understanding the functional diversity of cell wall mycolic acids of *Mycobacterium tuberculosis*. *Prog. Lipid Res.* 51, 325–339. 1873-2194 (Electronic). doi:10.1016/j.plipres.2012.05.002

Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., et al. (2013). Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* 13 (2), 137–146. doi:10.1016/S1473-3099(12)70277-3

Walker, T. M., Lalor, M. K., Broda, A., Ortega, L. S., Morgan, M., Parker, L., et al. (2014). Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir. Med.* 2 (4), 285–292. doi:10.1016/S2213-2600(14)70027-X

Weissman, K. J. (2006). Single amino acid substitutions alter the efficiency of docking in modular polyketide biosynthesis. *Chembiochem* 7 (1439-4227), 1334–1342. (Print)). doi:10.1002/cbic.200600185

Wilson, R., Kumar, P., Parashar, V., Vilchèze, C., Veyron-Churlet, R., Freundlich, J. S., et al. (2013). Antituberculosis thiophenes define a requirement for Pks13 in mycolic acid biosynthesis. *Nat. Chem. Biol.* 9 (1552-4469), 499–506. (Electronic)). doi:10.1038/nchembio.1277

Xu, Q., Zheng, J., Xu, R., Barany, G., and Cowburn, D. (1999). Flexibility of interdomain contacts revealed by topological isomers of bivalent consolidated ligands to the dual Src homology domain SH(32) of abelson. *Biochemistry* 38 (0006-2960), 3491–3497. (Print)). doi:10.1021/bi982744j

Yang, C., Lu, L., Warren, J. L., Wu, J., Jiang, Q., Zuo, T., et al. (2018). Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, genomic analysis. *Lancet Infect. Dis.* 18 (7), 788–795. doi:10.1016/S1473-3099(18)30218-4

Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., et al. (2017a). Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect. Dis.* 17, 275–284. 1474-4457 (Electronic). doi:10.1016/S1473-3099(16)30418-2

Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., et al. (2017b). Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect. Dis.* 17 (3), 275–284. doi:10.1016/S1473-3099(16)30418-2

Yu, J., Tran, V., Li, M., Huang, X., Niu, C., Wang, D., et al. (2012). Both phthiocerol dimycocerosates and phenolic glycolipids are required for virulence of Mycobacterium marinum. *Infect. Immun.* 80 (1098-5522), 1381–1389. (Electronic)). doi:10.1128/IAI.06370-11

Zelner, J. L., Murray, M. B., Becerra, M. C., Galea, J., Lecca, L., Calderon, R., et al. (2016). Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *J. Infect. Dis.* 213 (2), 287–294. doi:10.1093/infdis/jiv387

Zhang, H., Li, D., Zhao, L., Fleming, J., Lin, N., Wang, T., et al. (2013). Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* 45 (10), 1255–1260. doi:10.1038/ng.2735

TYPE Original Research
PUBLISHED 17 January 2024
DOI 10.3389/fmicb.2024.1343227

Check for updates

OPEN ACCESS

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Imchang Lee,
Hallym University, Republic of Korea
Sankarasubramanian Jagadesan,
University of Nebraska Medical Center,
United States

*CORRESPONDENCE
Ting Ma
✉ tingma@nankai.edu.cn

†These authors have contributed equally to this work

RECEIVED 23 November 2023
ACCEPTED 08 January 2024
PUBLISHED 17 January 2024

CITATION
Chen P, Wang S, Li H, Qi X, Hou Y and Ma T (2024) Comparative genomic analyses of *Cutibacterium granulosum* provide insights into genomic diversity.
Front. Microbiol. 15:1343227.
doi: 10.3389/fmicb.2024.1343227

Comparative genomic analyses of *Cutibacterium granulosum* provide insights into genomic diversity

Peishan Chen[1†], Shaojing Wang[2†], Hongyan Li[3,4†], Xiaoye Qi[3,4], Yuanyuan Hou[5] and Ting Ma[2*]

[1]Institute of Integrative Medicine for Acute Abdominal Diseases, Tianjin Nankai Hospital, Tianjin, China, [2]College of Life Sciences, Nankai University, Tianjin, China, [3]College of Bioengineering, Tianjin University of Science and Technology, Tianjin, China, [4]Tianjin JOYSTAR Technology Co., Ltd, Tianjin, China, [5]College of Pharmacy, Nankai University, Tianjin, China

Cutibacterium granulosum, a commensal bacterium found on human skin, formerly known as *Propionibacterium granulosum*, rarely causes infections and is generally considered non-pathogenic. Recent research has revealed the transferability of the multidrug-resistant plasmid pTZC1 between *C. granulosum* and *Cutibacterium acnes*, the latter being an opportunistic pathogen in surgical site infections. However, there is a noticeable lack of research on the genome of *C. granulosum*, and the genetic landscape of this species remains largely uncharted. We investigated the genomic features and evolutionary structure of *C. granulosum* by analyzing a total of 30 Metagenome-Assembled Genomes (MAGs) and isolate genomes retrieved from public databases, as well as those generated in this study. A pan-genome of 6,077 genes was identified for *C. granulosum*. Remarkably, the 'cloud genes' constituted 62.38% of the pan-genome. Genes associated with mobilome: prophages, transposons [X], defense mechanisms [V] and replication, recombination and repair [L] were enriched in the cloud genome. Phylogenomic analysis revealed two distinct mono-clades, highlighting the genomic diversity of *C. granulosum*. The genomic diversity was further confirmed by the distribution of Average Nucleotide Identity (ANI) values. The functional profiles analysis of *C. granulosum* unveiled a wide range of potential Antibiotic Resistance Genes (ARGs) and virulence factors, suggesting its potential tolerance to various environmental challenges. Subtype I-E of the CRISPR-Cas system was the most abundant in these genomes, a feature also detected in *C. acnes* genomes. Given the widespread distribution of *C. granulosum* strains within skin microbiome, our findings make a substantial contribution to our broader understanding of the genetic diversity, which may open new avenues for investigating the mechanisms and treatment of conditions such as acne vulgaris.

KEYWORDS

***C. granulosum*, genomics, phylogenomic, antibiotic resistance, horizontal gene transfer (HGT)**

1 Introduction

Acne is a persistent inflammatory skin condition, and its development is intricately linked to factors such as heightened sebum production, excessive skin keratinization, and bacterial overgrowth (Schommer and Gallo, 2013; Zaenglein et al., 2016). To treat acne, a range of antibiotics, including clindamycin, nadifloxacin, ozenoxacin, doxycycline, minocycline, and roxithromycin, are commonly employed (Hayashi et al., 2018). Nevertheless, the presence of antibiotic-resistant bacteria on skin poses a substantial threat to the efficacy of antibiotic treatments. Notably, bacteria carrying the multidrug-resistant plasmid pTZC1, which contains *erm* and *tet* genes responsible for resistance to macrolide-clindamycin and tetracycline, frequently exhibit antibiotic resistance (Nakase et al., 2012, 2020; Aoki et al., 2019, 2020; Koizumi et al., 2022). Consequently, the effective use of antibiotics is intimately connected to the emergence of antibiotic resistance in various skin commensal bacteria, including *Cutibacterium granulosum* (Koizumi et al., 2023b).

C. granulosum, formerly known as *Propionibacterium granulosum*, is a gram-positive bacterial species. It belongs to the *Cutibacterium* genus, which encompasses various bacterial species commonly found on human skin. As a part of skin microbiota, *C. granulosum* is primarily located in sebum-rich areas, although it is present at a much lower abundance than *C. acnes* (Park et al., 2020). The coexistence of *C. granulosum* and *C. acnes* has been observed, particularly in acne pustules (Koizumi et al., 2023b). *C. granulosum* exhibits high resistance rates to various antimicrobial agents, potentially posing challenges in treatment if it becomes the causative pathogen of opportunistic infections (Koizumi et al., 2023a). Recent comparative genomic studies have highlighted the transferability of the multidrug-resistant plasmid pTZC1 between *C. granulosum* and *C. acnes*, leading to the accumulation of Antibiotic Resistance Genes (ARGs) and an increased prevalence of multidrug-resistant strains on the skin surface (Koizumi et al., 2023b).

The utilization of Metagenome-Assembled Genomes (MAGs) in in pan-genomic analysis has become increasingly popular in recent years. Tang Li and Yanbin Yin conducted an evaluation of the use of MAGs in pan-genomics analysis by comparing the results of pan-genomic analysis between complete bacterial genomes and simulated MAGs (Li and Yin, 2022). Their findings addressed the impact of incompleteness and contamination on pan-genomic analysis (i.e., incompleteness led to core gene loss, while the contamination had influence on accessory genomes). They recommend utilizing higher quality MAGs, lowering core gene threshold, and employing metagenome mode for gene prediction. Most recently, Marcele Laux and colleagues updated the pan-genome of *Raphidiopsis* by adding newly generated MAGs (Laux et al., 2023). As mentioned in the article, they carefully validated these MAGs by both Average Nucleotide Identity (ANI) and Average Amino acid Identity (AAI) prior to pan-genomic analysis. These studies suggest that high-quality MAGs can be utilized in pan-genomic analysis.

Previous studies have underscored the importance and urgency of conducting more in-depth research on *C. granulosum*. However, there remains a noticeable scarcity of investigations into the genome of *C. granulosum*, leaving its genetic landscape largely unexplored. Thus, the primary objectives of this study are to unravel the genomic diversity and pinpoint the specific biological functions of *C. granulosum* genome, which might improve the understanding of its adaptation in surroundings. To achieve this goal, we conducted comparative genomic analyses, utilizing a total of 30 *C. granulosum* genomes retrieved from both isolates and metagenomes (Supplementary Figure S1).

2 Materials and methods

2.1 Participant recruitment and sample collection

Participants with good skin condition were recruited at the Tianjin Nankai Hospital, Tianjin, China, between July 15th, 2022 and September 12th, 2022. Ethical approval for the study was received from Tianjin Nankai Hospital research ethics committee in July 2022 (NKYY_YXKT_IRB _2022_040_01). Written informed consents were obtained from all participants. A total of 18 subjects meeting the inclusion criteria were recruited.

The subjects were asked not to clean their skin or apply any lotions, perfumes, cosmetics and other substances to their skin for 24-h prior to the metagenomic sampling. Sterile nonfat cotton swabs were soaked in sterile normal saline, and 4 cm^2 facial areas of left and right cheeks were wiped for 30 s each. After sampling, the cotton swab heads were placed into sterilized tubes, frozen in liquid nitrogen, and stored at −80°C. The sampling procedure was repeated three times.

2.2 DNA extraction and metagenomic sequencing

Genomic DNA was extracted from all the samples within one-week using a Wizard Genomic DNA purification kit (Promega, Madison, WI, United States) according to the manufacturer's recommended protocol. The quantity and purity of the extracted DNA were determined using a NanoDrop™ 2000 spectrophotometer (Thermo-Fisher Scientific, Waltham, MA, United States). Sequencing libraries were prepared according to the Nextera XT DNA Library Preparation Kit (Illumina) protocol and pooled to 2 mM. Paired-end sequencing run was performed on NextSeq 500 Illumina platform using NextSeq 500 v2 kit with 150 cycles.

2.3 Metagenomic data analysis

Adapter removal and quality trimming of the generated raw data were performed using TrimGalore v0.6.0[1] with default parameters. Host contaminated reads were removed using HoCoRT (Rumbavicius et al., 2023). MAGs were recovered using the metaWRAP pipeline (Uritskiy et al., 2018). Briefly, the modules of "assembly," "Binning" and "Bin_refinement" were executed step by step with default parameters. The taxonomy lineages for these MAGs were assigned with the classify workflow in GTDB-Tk v.2.1.1 (Chaumeil et al., 2022) by the Genome Taxonomy Database (GTDB) release R207_v2 (Parks et al., 2022). The lineage_wf

[1] https://github.com/FelixKrueger/TrimGalore

workflow of CheckM v1.0.18 (Parks et al., 2015) was employed to evaluate the quality of metagenomic bins. Only those *C. granulosum* MAGs with completeness ≥90% and contamination <5% were considered "high quality" (Singleton et al., 2021; Jiang et al., 2022) and retained for downstream analyses.

The MAGs sequences of *C. granulosum* originally generated in this study have been deposited in GenBank under accession numbers JAWMSC000000000 to JAWMST000000000. The corresponding raw reads have also been deposited in GenBank under accession numbers SRR27396937 to SRR27396954.

2.4 Public resourced genomes for pan-genome analysis

For comparative analysis, we collected all available genomic sequences for *C. granulosum*, comprising both MAGs and isolate genomes, from the National Center for Biotechnology Information (NCBI) website[2] as of September 10, 2023. Following this, all the genomes, including those originally presented in this study and those obtained from public database, underwent evaluation using the "lineage_wf" workflow within CheckM v1.07 (Parks et al., 2015). To standardize the genome assembly quality and minimize potential discrepancies in pan-genome analysis, only those genomes meeting the criteria of near completeness (completeness ≥ 90%) and low contamination (<5%) (Singleton et al., 2021; Jiang et al., 2022) were retained for further analysis.

Consequently, a total of 30 genomes of *C. granulosum* were included in this study, comprising 18 MAGs originally generated in this study and 12 retrieved from the NCBI genome database.

2.5 Comparative genomic analyses

For all retained genomes of *C. granulosum*, gene predictions were performed using Prokka v1.13 (Seemann, 2014). The resulting GFF3 files served as input for Roary v3.11.2 (Page et al., 2015) to delineate the pan-genome, which comprises the hard-core, soft-core, shell, cloud, and unique genomes. The hard-core genome is defined as the set of genes shared by all 30 *C. granulosum* genomes. The soft-core genome consists of the genes retained by 29 out of the 30 genomes. The shell and cloud genomes encompass gene sets shared by 5–28 and 1–4 genomes, respectively. The unique genome contains the set of genes observed in only one of the genomes.

The phylogenetic inference of *C. granulosum* was reconstructed using the alignment of hard-core genome generated by Roary. To eliminate potential recombination regions, Gubbins v3.0.0 (Croucher et al., 2015) was employed with the following options: "-m 4 -b 4,000 --first-tree-builder fasttree". Subsequently, a Maximum Likelihood (ML) tree was constructed using RAxML-NG v.1.1 (Kozlov et al., 2019) with rapid bootstrap 1,000 and the best fitting model of "GTR + G4" determined by ModelTest-NG v0.1.7 (Darriba et al., 2020). The final tree was visualized using iTOL (Letunic and Bork, 2021).

2.6 Comparative phylogenetic analyses

The 16S rRNA sequences of the 30 *C. granulosum* genomes were identified using Barrnap (0.9-dev, https://github.com/tseemann/barrnap). Multiple Sequence Alignment (MSA) and 16S rRNA-based phylogenetic tree based on ML algorithm were generated by MEGAX software (Kumar et al., 2018) with bootstrap 1,000.

Whole Genome Alignment (WGA) was constructed using the "nucmer" command of MUMmer (Marçais et al., 2018), with *C. granulosum* NCTC11865 (GCA_900186975.1) as the reference genome. MSA was constructed using whole-genome-wide Single Nucleotide Polymorphisms (SNPs) generated by the "show-snps" command, based on the coordinates of the reference genome. Potential recombination regions were eliminated using Gubbins v3.0.0 (Croucher et al., 2015). Subsequently, a WGA-based ML tree was constructed with RAxML-NG v.1.1 (Kozlov et al., 2019), employing rapid bootstrap (1,000 replicates) and the best-fitting model "GTR + G4" determined by ModelTest-NG v0.1.7 (Darriba et al., 2020).

Whole genome SNPs were identified using kSNP4 (Hall and Nisbet, 2023) with the parameters "-k 17 -annotate annotatedGenomes -ML -vcf." The kSNP4 tool employs an alignment-free approach for SNP identification. The SNPs-based ML tree was generated using FastTree (Price et al., 2010), which was automatically applied in the kSNP4 pipeline.

All phylogenetic trees were visualized using iTOL (Letunic and Bork, 2021).

2.7 ANI and AAI

ANI values between all genomes were calculated using fastANI v1.3 (Jain et al., 2018) with the parameter "--fragLen 100." The size of orthologous regions for ANI calculation were determined by the summation of orthologous matches multiplied by fragLen (100 bp). The pairwise ANI values were plotted using the ggplots package in R.[3] The AAI values and the proportion of matched CDS were calculated by EzAAI tool with default parameters (Kim et al., 2021).

2.8 Functional analyses

The distribution of Clusters of Orthologous Groups (COG) was automated using COGclassifier.[4]

The identification of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) regions and CRISPR-associated (Cas) proteins was conducted with CRISPRCasFinder (Couvin et al., 2018), using the genome sequences of *C. granulosum* as input.

The detection of ARGs was performed using the Comprehensive Antibiotic Resistance Database (CARD) due to its consistent maintenance and regular updates (Alcock et al., 2023). Protein sequences of the coding genes predicted in 30 *C. granulosum* genomes were annotated through a BLASTP search

[2] www.ncbi.nlm.nih.gov

[3] www.r-project.org

[4] https://github.com/moshi4/COGclassifier

against CARD 2023 (Alcock et al., 2023). The search parameters included E-value threshold of less than 1e-5 and minimum alignment length percentage of greater than 40%, as previously described by Zhang et al. (2019).

Putative virulence-related factors within the genomes of *C. granulosum* were identified by searching against the Virulence Factor Database (VFDB) (Liu et al., 2022). Each proteome was individually aligned with the VFDB full dataset using the BLASTp algorithm. A matrix was created based on VFDB hits against proteins in each genome. The matrix was filtered using BLASTp score threshold of ≥80 as described by Rasheed et al. (2017).

2.9 Horizontal gene transfer (HGT) analyses

MetaCHIP (Song et al., 2019) was employed to detect HGT among the 30 *C. granulosum* genomes with default parameters. A customized grouping file was provided to MetaCHIP, containing clade-specific information for each genome marked as "clade A," "clade B," or "others." The term "others" represents for the species not located in clade A nor in clade B.

3 Results

3.1 Genomic overview of *C. granulosum*

This study encompassed a total of 30 *C. granulosum* genomes, consisting of five genomes from pure cultured isolates and the remaining 25 genomes retrieved from metagenomes (Supplementary Table S1). The average genome size of *C. granulosum* is approximately 2.153 ± 0.103 Mb, with an average G + C content of 64.143% ± 0.289% and an average of 1860 ± 94 coding genes per genome. Notably, the genome with the ID GCA_032510525.1, retrieved from metagenomes of infant feces, exhibited a relatively low G + C content of 62.721% compared to the others. All genomes retained in this study demonstrate high assembly quality, with an average completeness of 97.298% ± 2.393% and an average contamination rate of 1.809% ± 1.422%. The average number of contigs for all 25 MAGs is 191, with an average N50 of 72.18 kb.

The distribution of ANI and AAI of MAGs against complete genomes were investigated using both *C. granulosum* TP-CG7 and *C. granulosum* NCTC11865 as references. The average ANI value of 25 MAGs against *C. granulosum* NCTC11865 was 97.35%, with orthologous regions accounting for an average of 88.73% of the *C. granulosum* NCTC11865 genome. For the 25 MAGs against *C. granulosum* TP-CG7, the average ANI value was 98.03%, and the orthologous regions accounted for an average of 89.27% of the *C. granulosum* TP-CG7 genome (Supplementary Table S2). The average AAI value of 25 MAGs against *C. granulosum* NCTC11865 was 97.60%, with matched CDS accounting for an average of 84.49% of the *C. granulosum* NCTC11865 proteome. For the 25 MAGs against *C. granulosum* TP-CG7, the average AAI value was 98.12%, and the matched CDS accounted for an average of 83.96% of the *C. granulosum* TP-CG7 proteome (Supplementary Table S3). Both ANI and AAI values exceed 95%, with a significant portion of the genome/proteome utilized.

3.2 Pan-genome determination and distribution of functional categories

In order to assess genomic conservation across the 30 *C. granulosum* genomes, we conducted a comprehensive pan-genome analysis for defining the sets of hard-core, soft-core, shell, cloud, and unique genomes. A total of 6,077 genes representing the pan-genome of *C. granulosum* were revealed in this study (Figure 1A). Notably, the pan genome did not reach a plateau, as the addition of each new genome continued to increase the number of genes in the pan-genome (Figure 1B).

Among the 6,077 identified coding genes (Supplementary Table S4), 526 genes were shared across all 30 genomes, categorizing them as "hard-core" genes, while 412 genes were shared by at least 29 genomes, classifying them as 'soft-core' genes. The analysis also revealed 1,348 "shell" genes, which were shared by 5–28 genomes, and 3,791 "cloud" genes, shared by 1–4 genomes. Notably, the "cloud" genes constituted 62.38% of the pan-genome. It was observed that approximately 40% (2,443 out of 6,077) of the pan-genome consisted of "unique" genes, observed in only one of these genomes. The presence of "shell" and "cloud" genomes, especially the "unique" genome, contributes significantly to the genetic diversity and, presumably, phenotypic differences among strains.

Functional categories were assigned using COGclassifier, a tool designed for classifying prokaryotic protein sequences into COG functional categories. The number of genes for each functional category was represented with bar graphs (Figure 1C). Three basic biological functional categories were enriched in core-genome (union of hard- and soft-core genomes), i.e., nucleotide transport and metabolism [F], energy production and conversion [C] and translation, ribosomal structure and biogenesis [J]. Another three functional categories of mobilome: prophages, transposons [X], defense mechanisms [V] and replication, recombination and repair [L] were enriched in the cloud genome. The "cloud" genome is the set of genes shared by four genomes or fewer, making up a significant portion (62.38%) of the pan-genome for *C. granulosum*.

3.3 Phylogenomic analysis

To investigate the relationships between the *C. granulosum* genomes generated in this study and those available in public database, the core genome was used to reconstruct the phylogenetic tree. The most striking outcome of the phylogenomic analysis was the highly diverse relationships observed within *C. granulosum*. As depicted in Figure 2, the phylogenetic tree revealed two distinct monophyletic clades. Clade A consisted of ten genomes originally retrieved from skin surface of adult volunteers living in Tianjin, China. Clade B was comprised of seven genomes retrieved from isolates ($n = 4$, type materials) and metagenomes ($n = 3$, samples of infant feces collected at Magee-Womens Hospital of UPMC, Pittsburgh, United States).

The 16S rRNA sequences were identified in all 5 isolate genomes and only one MAG (Supplementary Figure S2). Owning to the miss identification of 16S rRNA in mostly MAGs, the identification of clade A and clade B was verified using WGA-based phylogenetic tree (Supplementary Figure S3) and SNPs-based phylogenetic tree (Supplementary Figure S4). Consistent monophyletic clades A and B, representing the same genomes, were observed in the phylogenetic trees based on WGA, SNPs and core genome (Figure 2; Supplementary Figures S3, S4).

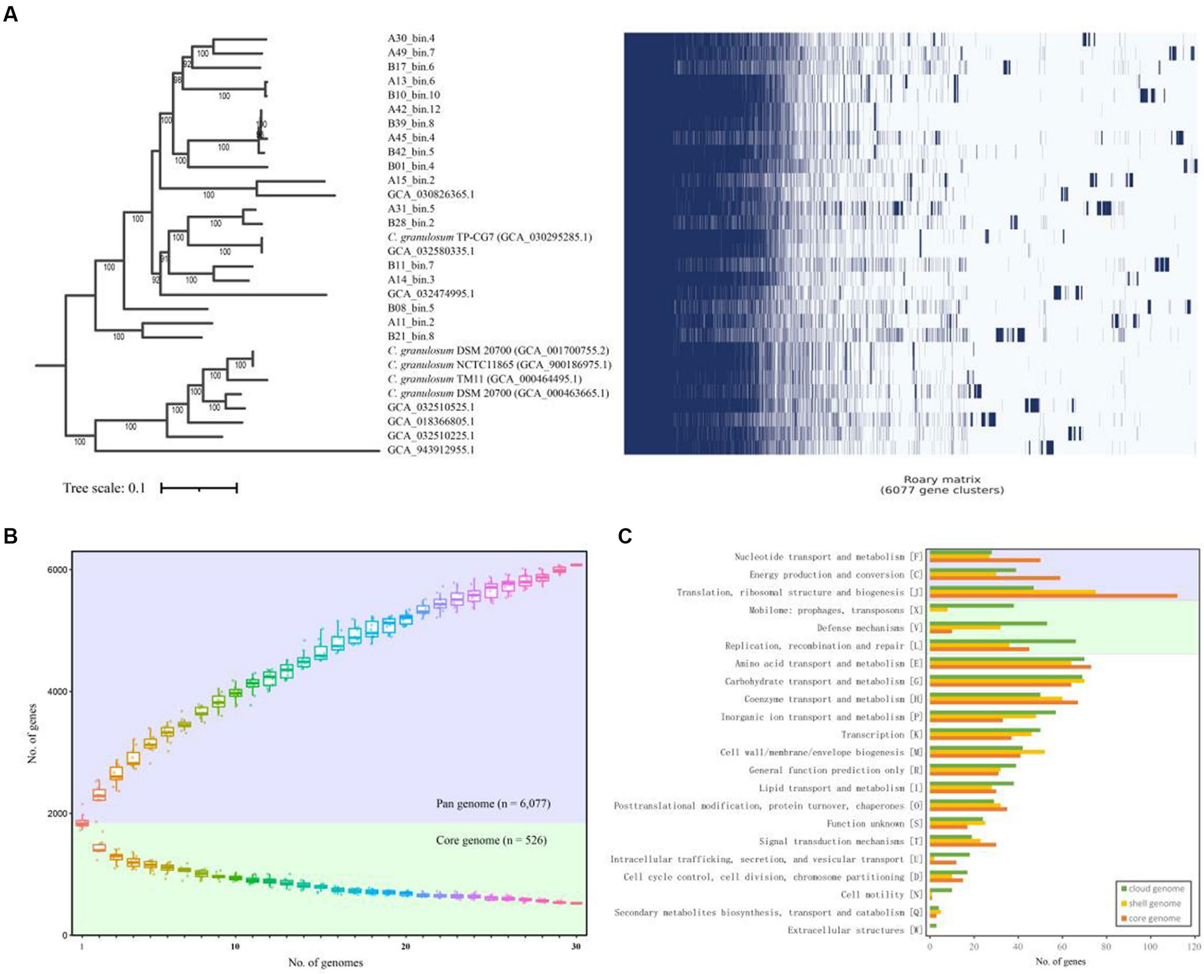


FIGURE 1
Pan- and core-genome of *C. granulosum*. **(A)** Gene cluster matrix of presence/absence (blue)/(white) of the 6,077 genes (columns) that constitute the pan genome of *C. granulosum* (rows). **(B)** The number of genes increases in the pan genome (blue background) and decreases in the core genome (green background) with the addition of more genomes. **(C)** Number of genes in the core (hard- and soft-core genes), shell and cloud genomes assigned to each functional COG category.

The cladistic relationships and genomic diversity were further validated using the ANI approach, a robust measure of genomic relatedness between strains (Figures 2, 3A). The overall ANI values across *C. granulosum* averaged at 97.45% ± 0.74%. Notably, the intra-clade ANI values of genomes in Clade A and Clade B were 98.55% ± 0.61 and 98.47% ± 0.42%, respectively. However, the inter-clade ANI values between genomes in Clade A and B were notably lower, measuring only 96.91% ± 0.16%. These significant differences in ANI values between intra-clade and inter-clade comparisons (Figure 3A, $p < 0.0001$) highlight the genetic diversity among these strains. Although no differences were observed in other genomic features such as genome size, G + C content, and the number of protein-coding genes (Figures 3B–D), the distant relationships between *C. granulosum* isolates and the larger proportion of cloud genome within the pan-genome offer a valuable perspective on the acquisition of biological functions.

A total of 64 interclade HGT events were observed (Supplementary Table S5), including 48 HGT events between species in clade A and others, and 16 HGT events between species in clade B and others. No HGT event was observed between species located in "clade A" and "clade B." Besides, a total of 1,463 clade-specific genes were observed, comprising 799 genes identified only in strains within Clade A and 664 genes identified only in strains within Clade B. The comparation of COG categories distribution for clade-specific genes was showed as Supplementary Figure S5. The results revealed that specific genes in Clade A were enriched in the categories of "[L] Replication, recombination and repair," "[X] Mobilome: prophages, transposons," "[C] Energy production and conversion" and "[O] Posttranslational modification, protein turnover, chaperones," compared to those in Clade B (two-fold change). Meanwhile, specific genes in Clade B were enriched in the categories of "[I] Lipid transport and metabolism," "[G] Carbohydrate transport and metabolism," "[J] Translation, ribosomal structure and biogenesis," "[V] Defense mechanisms." Additionally, a total of 1,284 clade A-specific SNPs and 4,292 clade B-specific SNPs were identified (Supplementary Table S6).

3.4 Functional analysis of *C. Granulosum* genomes

To gain insights into the functional profiles of *C. granulosum*, the distributions of putative ARGs, virulence-related factors, and CRISPR-Cas systems were investigated.

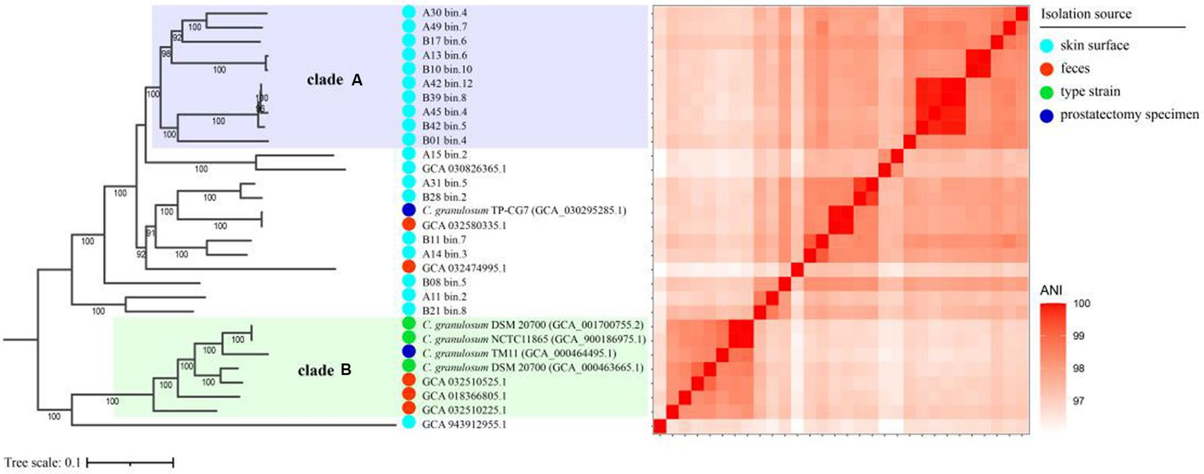


FIGURE 2
Phylogenomic analysis of *C. granulosum*. The phylogenetic tree reconstructed using core genome. The heatmap pairwise the distribution of pairwise ANI values across the phylogenetic tree.

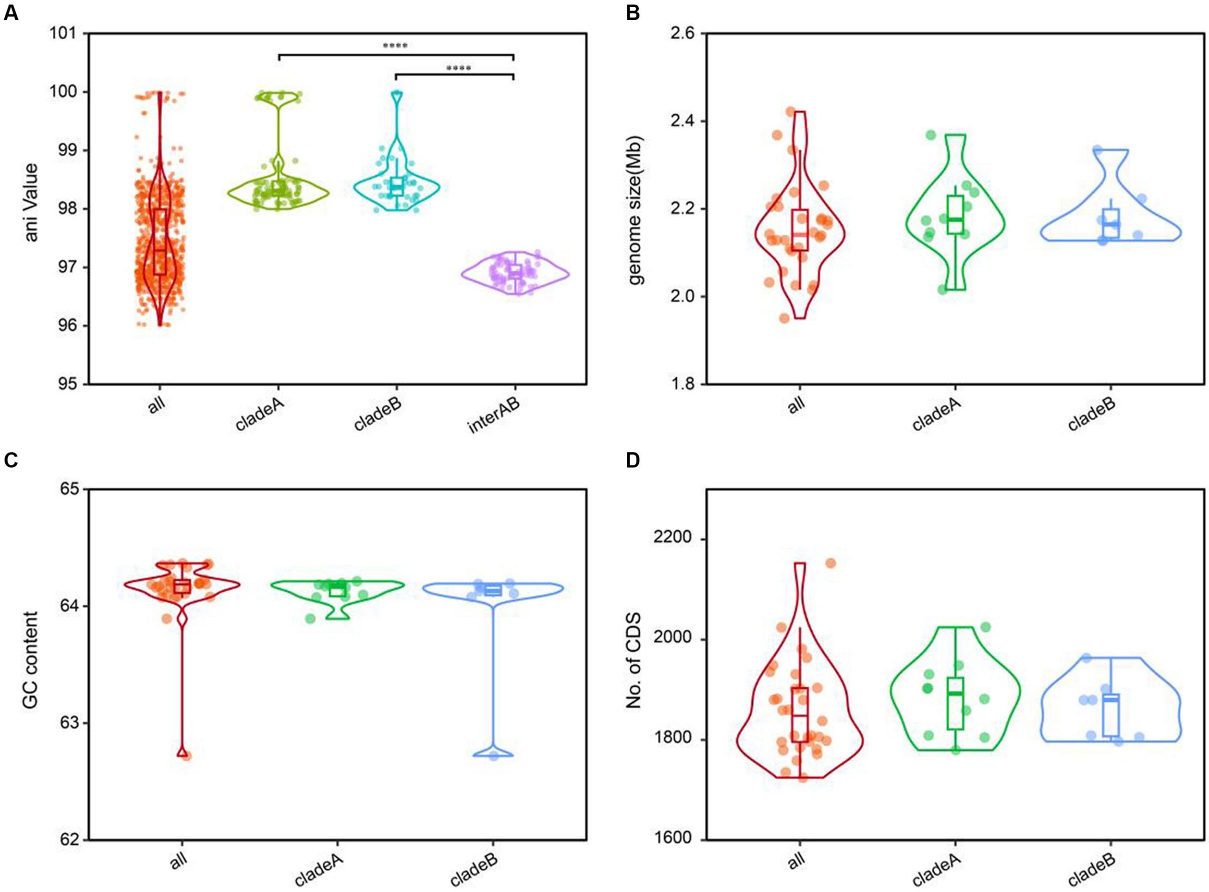


FIGURE 3
Comparative genomic features of *C. granulosum*. **(A)** Comparation of pairwise ANI values. **(B)** Comparation of genome size. **(C)** Comparation of genomic G+C content. **(D)** Comparation of protein coding genes number.

The ARGs were identified through a BLASTp search against the CARD database. The CARD database is a meticulously curated resource comprising antibiotics, their targets, ARGs, associated proteins, and antibiotic resistance literature. The analysis revealed a wide range of potential ARGs within the *C. granulosum* genomes, suggesting its potential tolerance to various environmental challenges (Figure 4A; Supplementary Table S7).

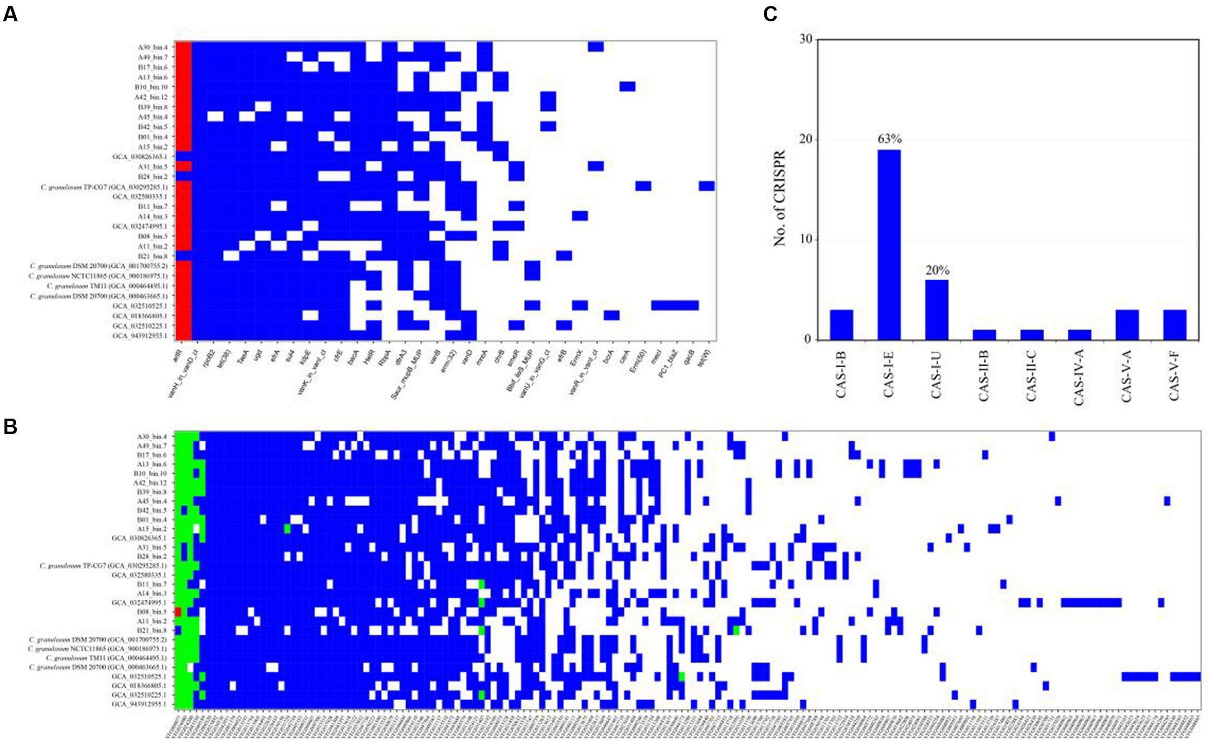


FIGURE 4
Distribution of functional genes in *C. granulosum* genomes. **(A)** Distribution of ARGs based on CARD. Blue markings denote the existence of ARGs within the corresponding genome, whereas red dots indicate the presence of two copies of the respective ARGs within the genome. **(B)** Distribution of virulence factors based on VFDB. Blue markings denote the existence of presumed virulence factors, whereas green and red dots indicate the presence of two or three copies of presumed virulence factors within the genome, respectively. **(C)** Occurrence and diversity of CRISPR-Cas systems.

Putative virulence-related factors were identified based on the VFDB database, known for its comprehensive coverage of virulence factors and detailed information, including structural features, functions, and mechanisms. Our analysis revealed approximately 72 ± 8 genes encoding for virulence-related factors in each *C. granulosum* genome (Figure 4B; Supplementary Table S8).

CRISPR-Cas systems constitute the bacterial adaptive immune system, providing resistance against bacteriophage infections (Barrangou et al., 2007). This system is known for its adaptability and heritability. In our study, we investigated the presence and diversity of CRISPR-Cas systems across the 30 *C. granulosum* genomes. Subtype I-E of the CRISPR-Cas system was identified in 63% (19 out of 30) of the genomes (Figure 4C), based on the presence of the corresponding signature *cas* genes.

4 Discussion

Cutibacterium strains, particularly *C. granulosum* and *C. acnes*, are crucial members of human skin microbiome and play vital roles in both skin health and disease (Cobian et al., 2021; Koizumi et al., 2023b). Recently, the study by Juri Koizumi and colleagues emphasized the significance of *C. granulosum* in skin health and disease by demonstrating the horizontal transfer of the plasmid pTZC1 between *C. acnes* and *C. granulosum* strains (Koizumi et al., 2023b). This multidrug resistance plasmid pTZC1, which carries genes for macrolide-clindamycin resistance (*erm*(50)) and tetracycline resistance (*tet*(W)), can lead to antimicrobial resistance.

C. acnes is the most extensively studied species, and a substantial number of corresponding genomes have been sequenced and are accessible in public databases. Currently, the NCBI genome database[5] contains a total of 497 genomes of *C. acnes*. A previous study by Natalia Cobian and colleagues involved a comparative genomic analysis of 255 *C. acnes* genomes (Cobian et al., 2021). These results revealed the genomic landscape of *C. acnes* including pan-genome, distinct phylogenetic clades and diverse CRISPR-Cas systems.

In this study, we present a comparative genomic analysis of *C. granulosum*, utilizing MAGs and genomes of isolated strains. The pan-genome depiction of gene presence and absence unveiled accessory genes and gene groups contributing to *C. granulosum*'s functional diversity. Our findings revealed a relatively small and open pan-genome comprising 6,077 genes, with a substantial "cloud" genome accounting for 62.38% of the pan-genome. Remarkably, for the first time, we observed two distinct phylogenetic clades representing strains retrieved from distant environments, a pattern somewhat similar to the genomic diversity found in *C. acnes* (Cobian et al., 2021). Our research significantly augments the dataset of *C. granulosum* genomes and provides an encompassing genomic landscape through pan-genome analysis and functional assessment.

Based on the gene distribution across COG categories, we could infer the genetic diversity of *C. granulosum* in terms of transposons, recombination, and defense mechanisms. While it remains challenging

[5] www.ncbi.nlm.nih.gov/genome

to directly associate these clades with specific ecological niches due to the limited number of genomes available at present, the results presented here highlight the genomic diversity of *C. granulosum*. The data provided in this study significantly contributes to the understanding of this species and expands the research foundation for future studies. The prevalence of CRISPR-Cas systems subtypes I-E in *C. granulosum*, as also detected in *C. acnes* genomes (Cobian et al., 2021), offers insights into likely shared strain divergence and adaptive differentiation between these two species. Furthermore, our results serve as a foundational reference and present new prospects for modulating the composition of skin microbiota using naturally occurring phages, engineered phages, and/or heterologous CRISPR-Cas systems.

The knowledge presented here is essential for understanding the role of *C. granulosum* in skin ecosystem and its potential applications, including the possibilities of utilizing *C. granulosum* to maintain skin health, advance biotechnological applications, and foster innovation in the fields of cosmetics and pharmaceuticals. Future research endeavors will continue to unveil the precise role of *C. granulosum* in promoting skin health and maintaining microbial balance, thus accelerating developments in its various application areas. While this study has made a significant contribution by substantially expanding the number of *C. granulosum* genomes and including high-quality accessible genomes from public databases, it is worth noting that the overall available genomic data for this species remains relatively limited.

5 Conclusion

Given the widespread distribution of diverse *C. granulosum* within skin microbiome, whole-genome sequencing offers valuable insights into its roles in health and disease. Comparative genomics analyses provide a robust method for examining extensive genome datasets. Our findings significantly contribute to the broader understanding of the genetic diversity within *C. granulosum*. Notably, our study is the first to reveal the presence of two distinct phylogenetic clades based on genomic data. Understanding the differential genetic content among *C. granulosum* strains in future research may open new avenues for investigating the mechanisms and treatment of conditions such as acne vulgaris.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

PC: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. SW: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. HL: Data curation, Validation, Writing – review & editing. XQ: Formal analysis, Writing – review & editing. YH: Formal analysis, Writing – review & editing. TM: Formal analysis, Methodology, Resources, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Guangxi Innovation-driven Development Special Foundation Project (No. AA22117003-4), the Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (No. TSBICIP-KJGG-015-04), the National Natural Science Foundation of China (No. 32300098), and the China Postdoctoral Science Foundation (No. 2023M731808).

Conflict of interest

HL and XQ were employed by Tianjin JOYSTAR Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1343227/full#supplementary-material

References

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., et al. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 51, D690–D699. doi: 10.1093/nar/gkac920

Aoki, S., Nakase, K., Hayashi, N., and Noguchi, N. (2019). Transconjugation of erm(X) conferring high-level resistance of clindamycin for Cutibacterium acnes. *J. Med. Microbiol.* 68, 26–30. doi: 10.1099/jmm.0.000875

Aoki, S., Nakase, K., Nakaminami, H., Wajima, T., Hayashi, N., and Noguchi, N. (2020). Transferable multidrug-resistance plasmid carrying a novel macrolide-clindamycin resistance gene, erm(50), in Cutibacterium acnes. *Antimicrob. Agents Chemother.* 64, e01810–e01819. doi: 10.1128/AAC.01810-19

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712. doi: 10.1126/science.1138140

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38, 5315–5316. doi: 10.1093/bioinformatics/btac672

Cobian, N., Garlet, A., Hidalgo-Cantabrana, C., and Barrangou, R. (2021). Comparative genomic analyses and CRISPR-Cas characterization of Cutibacterium acnes provide insights into genetic diversity and typing applications. *Front. Microbiol.* 12:758749. doi: 10.3389/fmicb.2021.758749

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., et al. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* 46, W246–W251. doi: 10.1093/nar/gky425

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Res.* 43:e15. doi: 10.1093/nar/gku1196

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294. doi: 10.1093/molbev/msz189

Hall, B. G., and Nisbet, J. (2023). Building phylogenetic trees from genome sequences with kSNP4. *Mol. Biol. Evol.* 40:msad235. doi: 10.1093/molbev/msad235

Hayashi, N., Akamatsu, H., Iwatsuki, K., Shimada-Omori, R., Kaminaka, C., Kurokawa, I., et al. (2018). Japanese dermatological association guidelines: guidelines for the treatment of acne vulgaris 2017. *J. Dermatol.* 45, 898–935. doi: 10.1111/1346-8138.14355

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9

Jiang, F., Li, Q., Wang, S., Shen, T., Wang, H., Wang, A., et al. (2022). Recovery of metagenome-assembled microbial genomes from a full-scale biogas plant of food waste by pacific biosciences high-fidelity sequencing. *Front. Microbiol.* 13:1095497. doi: 10.3389/fmicb.2022.1095497

Kim, D., Park, S., and Chun, J. (2021). Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *J. Microbiol.* 59, 476–480. doi: 10.1007/s12275-021-1154-0

Koizumi, J., Nakase, K., Hayashi, N., Nasu, Y., Hirai, Y., and Nakaminami, H. (2023a). Prevalence of antimicrobial-resistant Cutibacterium isolates and development of multiplex PCR method for Cutibacterium species identification. *J. Infect. Chemother.* 29, 198–204. doi: 10.1016/j.jiac.2022.10.018

Koizumi, J., Nakase, K., Hayashi, N., Takeo, C., and Nakaminami, H. (2023b). Multidrug resistance plasmid pTZC1 could be pooled among Cutibacterium strains on the skin surface. *Microbiol. Spectr.* 11:e0362822. doi: 10.1128/spectrum.03628-22

Koizumi, J., Nakase, K., and Nakaminami, H. (2022). Identification of a transferable linear plasmid carrying the macrolide-clindamycin resistance gene erm(X) in a Cutibacterium acnes isolate from a patient with acne vulgaris in Japan. *Microbiol. Resour. Announc.* 11:e0009422. doi: 10.1128/mra.00094-22

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Laux, M., Piroupo, C. M., Setubal, J. C., and Giani, A. (2023). The Raphidiopsis (= Cylindrospermopsis) raciborskii pangenome updated: two new metagenome-assembled genomes from the South American clade. *Harmful Algae* 129:102518. doi: 10.1016/j.hal.2023.102518

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Li, T., and Yin, Y. (2022). Critical assessment of pan-genomic analysis of metagenome-assembled genomes. *Brief. Bioinform.* 23:bbac413. doi: 10.1093/bib/bbac413

Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50, D912–D917. doi: 10.1093/nar/gkab1107

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14:e1005944. doi: 10.1371/journal.pcbi.1005944

Nakase, K., Aoki, S., Sei, S., Fukumoto, S., Horiuchi, Y., Yasuda, T., et al. (2020). Characterization of acne patients carrying clindamycin-resistant Cutibacterium acnes: a Japanese multicenter study. *J. Dermatol.* 47, 863–869. doi: 10.1111/1346-8138.15397

Nakase, K., Nakaminami, H., Noguchi, N., Nishijima, S., and Sasatsu, M. (2012). First report of high levels of clindamycin-resistant Propionibacterium acnes carrying erm(X) in Japanese patients with acne vulgaris. *J. Dermatol.* 39, 794–796. doi: 10.1111/j.1346-8138.2011.01423.x

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Park, S. Y., Kim, H. S., Lee, S. H., and Kim, S. (2020). Characterization and analysis of the skin microbiota in acne: impact of systemic antibiotics. *J. Clin. Med.* 9:168. doi: 10.3390/jcm9010168

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P. A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 50, D785–D794. doi: 10.1093/nar/gkab776

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

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rasheed, M. A., Qi, J., Zhu, X., Chenfei, H., Menghwar, H., Khan, F. A., et al. (2017). Comparative genomics of mycoplasma bovis strains reveals that decreased virulence with increasing passages might correlate with potential virulence-related factors. *Front. Cell. Infect. Microbiol.* 7:177. doi: 10.3389/fcimb.2017.00177

Rumbavicius, I., Rounge, T. B., and Rognes, T. (2023). HoCoRT: host contamination removal tool. *BMC Bioinformatics* 24:371. doi: 10.1186/s12859-023-05492-w

Schommer, N. N., and Gallo, R. L. (2013). Structure and function of the human skin microbiome. *Trends Microbiol.* 21, 660–668. doi: 10.1016/j.tim.2013.10.001

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

Singleton, C. M., Petriglieri, F., Kristensen, J. M., Kirkegaard, R. H., Michaelsen, T. Y., Andersen, M. H., et al. (2021). Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* 12:2009. doi: 10.1038/s41467-021-22203-2

Song, W., Wemheuer, B., Zhang, S., Steensen, K., and Thomas, T. (2019). MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* 7:36. doi: 10.1186/s40168-019-0649-y

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1

Zaenglein, A. L., Pathy, A. L., Schlosser, B. J., Alikhan, A., Baldwin, H. E., Berson, D. S., et al. (2016). Guidelines of care for the management of acne vulgaris. *J. Am. Acad. Dermatol.* 74, 945–973.e33. doi: 10.1016/j.jaad.2015.12.037

Zhang, X., Xiao, S., Jiang, X., Li, Y., Fan, Z., Yu, Y., et al. (2019). Genomic characterization of Escherichia coli LCT-EC001, an extremely multidrug-resistant strain with an amazing number of resistance genes. *Gut Pathog.* 11:25. doi: 10.1186/s13099-019-0298-5

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 24 January 2024
DOI 10.3389/fmicb.2023.1340427

Comparative genomic analysis of uropathogenic *Escherichia coli* strains from women with recurrent urinary tract infection

Marco A. Flores-Oropeza[1,2], Sara A. Ochoa[2], Ariadnna Cruz-Córdova[2], Rolando Chavez-Tepecano[3], Eva Martínez-Peñafiel[2], Daniel Rembao-Bojórquez[4], Sergio Zavala-Vega[4,5], Rigoberto Hernández-Castro[6], Marcos Flores-Encarnacion[7], José Arellano-Galindo[8], Daniel Vélez[9,10]* and Juan Xicohtencatl-Cortes[2]*

[1]Posgrado en Ciencias Biomédicas, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2]Laboratorio de Investigación en Bacteriología Intestinal, Unidad de Enfermedades Infecciosas, Hospital Infantil de México Federico Gómez, Mexico City, Mexico, [3]Ginecología del Hospital Militar Regional Especialidades de Monterrey, Monterrey, Mexico, [4]Departamento de Patología, Instituto Nacional de Neurología y Neurocirugía, Manuel Velasco Suárez, Mexico City, Mexico, [5]Laboratorio Clínico y Banco de Sangre, Instituto Nacional de Neurología y Neurocirugía, Manuel Velasco Suárez, Mexico City, Mexico, [6]Departmento de Ecología de Agentes Patógenos, Hospital General "Dr. Manuel Gea González", Mexico City, Mexico, [7]Laboratorio de Microbiología Molecular y Celular, Biomedicina, Facultad de Medicina, BUAP, Puebla, Mexico, [8]Laboratorio de Virología Clínica y Experimental, Unidad de Investigación en Enfermedades Infecciosas, Hospital Infantil de México Federico Gómez, Mexico City, Mexico, [9]Hospital Militar de Especialidades de la Mujer y Neonatología, Mexico City, Mexico, [10]Unidad Médica de Alta Especialidad, Hospital de Ginecología y Obstetricia No. 3 IMSS, Mexico City, Mexico

OPEN ACCESS

EDITED BY
Augusto Simoes-Barbosa,
The University of Auckland, New Zealand

REVIEWED BY
Payam Behzadi,
Islamic Azad University, ShahreQods, Iran
Lara Ferrero Gómez,
Universidade Jean Piaget de Cabo,
Cabo Verde

*CORRESPONDENCE
Daniel Vélez
✉ dr.danielvelez@gmail.com
Juan Xicohtencatl-Cortes
✉ juanxico@yahoo.com

RECEIVED 17 November 2023
ACCEPTED 21 December 2023
PUBLISHED 24 January 2024

CITATION
Flores-Oropeza MA, Ochoa SA, Cruz-Córdova A, Chavez-Tepecano R, Martínez-Peñafiel E, Rembao-Bojórquez D, Zavala-Vega S, Hernández-Castro R, Flores-Encarnacion M, Arellano-Galindo J, Vélez D and Xicohtencatl-Cortes J (2024) Comparative genomic analysis of uropathogenic *Escherichia coli* strains from women with recurrent urinary tract infection. *Front. Microbiol.* 14:1340427. doi: 10.3389/fmicb.2023.1340427

Introduction: Recurrent urinary tract infections (RUTIs) caused by uropathogenic *Escherichia coli* are costly public health problems impacting patients' quality of life.

Aim: In this work, a comparative genomics analysis of three clinical RUTI strains isolated from bladder biopsy specimens was performed.

Materials and methods: One hundred seventy-two whole genomes of urinary tract *E. coli* strains were selected from the NCBI database. The search for virulence factors, fitness genes, regions of interest, and genetic elements associated with resistance was manually carried out. The phenotypic characterization of antibiotic resistance, haemolysis, motility, and biofilm formation was performed. Moreover, adherence and invasion assays with human bladder HTB-5 cells, and transmission electron microscopy (TEM) were performed.

Results: The UTI-1_774U and UTI-3_455U/ST1193 strains were associated with the extraintestinal pathotypes, and the UTI-2_245U/ST295 strain was associated with the intestinal pathotype, according to a phylogenetic analysis of 172 *E. coli* urinary strains. The three RUTI strains were of clinical, epidemiological, and zoonotic relevance. Several resistance genes were found within the plasmids of these strains, and a multidrug resistance phenotype was revealed. Other virulence genes associated with CFT073 were not identified in the three RUTI strains (genes for type 1 and P fimbriae, haemolysin *hlyA,* and *sat* toxin). Quantitative adherence analysis showed that UTI-1_774U was significantly ($p < 0.0001$) more adherent to human bladder HTB-5 cells. Quantitative invasion analysis showed that UTI-2_245U was significantly more invasive than the control strains. No haemolysis or biofilm activity was detected in the three RUTI strains. The TEM micrographs showed the presence of short and thin fimbriae only in the UTI-2_245U strain.

Conclusion: The high variability and genetic diversity of the RUTI strains indicate that are a mosaic of virulence, resistance, and fitness genes that could promote recurrence in susceptible patients.

KEYWORDS

UPEC, RUTIs, MDR, WGS, adherence

1 Introduction

Recurrent urinary tract infections (RUTIs) are a public health problem in Mexico, mainly due to their difficulty in eradicating and the appearance of repeated infectious episodes with disabling signs and symptoms. RUTIs decrease the patient's quality of life and frequently cause the use of various antibiotic regimens, which can generate bacterial multidrug resistance (Ochoa et al., 2016). UPEC is the main uropathogen recovered from UTIs in vulnerable groups such as children (Contreras-Alvarado et al., 2021), pregnant women (Acuña-Ruíz and Molina-Torres, 2022), women of reproductive age (Ballesteros-Monrreal et al., 2021) and older adults. More than 4 million cases in Mexico have been reported yearly (Secretaría de Salud, 2023). The recurrence rate of UTIs in Mexico has been reported to be 20–23%, depending on age and study groups (Acuña-Ruíz and Molina-Torres, 2022).

Urinary tract infections (UTIs) are an inflammatory response to the colonization and multiplication of a microorganism in any organ of the urinary system. These infections have been associated with dysuria, haematuria, frequent urination, urgency, and occasionally suprapubic pain (Flores-Mireles et al., 2015; Anger et al., 2019; Burnett et al., 2021). The microbiological diagnosis of UTIs is based on the colony counts in a urine culture, and a bacterial concentration of 1 × 10^5 CFU/mL (colony forming units per milliliter) typically indicates a positive test result according to the Kass-Sandford criteria (Kass, 1956). UTIs occur more frequently in women; the incidence rate of cystitis is 12.6% in women and 3.0% in men in the United States (Wagenlehner et al., 2020). Other data have estimated that more than half of the female population suffer from a UTI episode in their lifetime. Moreover, 25% of women with UTIs experience recurrent infection 3 months after the initial episode (Stamm and Hooton, 1993; Foxman, 2010). Recurrent urinary tract infections (RUTIs) occur when a patient has more than two UTI episodes in 6 months or more than three UTI episodes in a year (Haddad et al., 2020). RUTIs represent a global health problem and can cause chronic infection, poor outcomes, and decreased patient quality of life (Wagenlehner et al., 2018). Recurrent cystitis in women is usually disabling, with 2.78 visits to the physician and 3.45 days of limited activity per year. Many women develop numerous symptoms, and more than 80.3% are treated with various antibiotics. Recurrence prophylaxis is a treatment option; however, the infection was still present in 73% of the patients treated with antibiotic prophylaxis, causing more mental stress in patients (Wagenlehner et al., 2018).

Approximately 80% of all RUTIs are caused by uropathogenic *Escherichia coli* (UPEC) (Thanert et al., 2019). However, the pathophysiology of RUTIs caused by UPEC is a complex process that has not yet been fully characterized. Several virulence factors (VFs) contribute to UPEC pathogenicity, such as characterized fimbriae (Auf, Dr., F1C, S, Type 9, Type 1, and P fimbria), un characterized fimbriae (Type 3 fimbriae), siderophores (enterobactin, aerobactin, and yersiniabactin), toxins (haemolysin, autotransporter secreted toxin, and autotransporter vacuolating toxin), capsule production and variations in lipopolysaccharide (Behzadi, 2020; Khonsari et al., 2021). The presence of metabolic regulators and protectins, increasing serum survival factor (*iss*) and proline permease (*pro*), has been related to the persistence of UPEC (Johnson and Russo, 2018). Multidrug-resistant (MDR) UPEC strains that cause RUTIs are a serious global problem as the antimicrobial treatments necessary to treat them are costly (Brown and Wright, 2016; Habibi and Khameneie, 2016). The discovery of genes encoding extended-spectrum beta-lactamases (ESBL), genes that endow resistance to quinolones, and sulfonamides has contributed vital information to clinical treatment guidelines in Mexico (Alcántar-Curiel et al., 2015). In addition, 5% of RUTIs-associated UPEC strains form intracellular bacterial communities (IBCs) and reactivate quiescent reservoirs (QRs) within the inner layers of the uroepithelium (Mulvey et al., 2001). RUTIs-causing UPEC uses mechanisms such as the production of resistance genes that can lead to treatment failure and the selection of MDR strains that persist in the urinary tract. Moreover, UPEC MDR strains can reactivate once antimicrobial treatment has been completed (Ulett and Schembri, 2016).

Activation of invasins, toxins, siderophores, and metabolic fitness factors contributes to the invasion of uroepithelial cells to form intracellular bacterial communities (IBCs). During this process, UPEC remains within the cells of the transitional urinary epithelium, isolated and protected from antimicrobial treatment and the host's immune response (Kim et al., 2021). QRs formed by UPEC in the deepest layers of the urinary epithelium can remain for long periods and be reactivated by various signaling systems, a mechanism that has not yet been described in detail; however, these mechanisms can be independent or cooperate to perform a single process (Mulvey et al., 2001).

UPEC strains can acquire mobile genetic elements (plasmids, prophages, or genomic islands) that play a role in horizontal gene transfer (Desvaux et al., 2020). Mobile genetic elements can carry virulence, resistance, and fitness genes that help the bacterium colonize various ecological niches (Baldy-Chudzik et al., 2015). Pathogenicity islands (PAIs) are genomic elements primarily associated with virulence genes that code for resistance determinants, colonization factors, nutrient acquisition, and toxins mediated by specialized secretion systems. PAIs are important elements that distinguish diarrhoeagenic and extraintestinal pathotypes of *E. coli* (Lloyd et al., 2007). Highly virulent strains carrying PAI-I$_{CFT073}$ have been found to carry genes encoding haemolysin, P fimbriae, iron-regulated gene homolog adhesion (*iha*), aerobactin, secreted autotransporter toxin (*sat*), *agn43*, and the "K" capsule widely associated with uropathogenesis (Schneider et al., 2004; Lloyd et al., 2007).

Several treatment options for RUTIs have been proposed, including continuous or postcoital antibiotic prophylaxis, behavioral therapy, treatment with probiotics, treatment with oestrogens, and intravesical instillations of hyaluronate. However, the results have been unsuccessful in their short-term and long-term efficacy (Pigrau and Escolà-Vergé, 2020). In this context, new approaches are needed to design prevention strategies for RUTIs and improve patient quality of life (Behzadi et al., 2023). Bacterial whole-genome sequencing (WGS) is the most recent and up-to-date technology for characterizing the whole genome of UPEC strains, allowing us to understand their genomic context (Voelkerding et al., 2009). This work performed complete genome sequencing of three clinical UPEC strains. The purpose was to conduct a comparison and characterization study on the genomes of UPEC strains isolated from adult women treated at the Military Hospital for Specialties for Women and Neonatology in Mexico City for RUTIs reoccurring over a year. In this study, potential VFs, aptitude factors, and antibiotic resistance were investigated, including the search for elements responsible for the establishment, persistence, and multidrug-resistance (MDRs) of RUTIs. Establishing profiles responsible for recurrence will aid the development of strategies for preventing and treating these infections.

2 Materials and methods

2.1 Wet lab section

2.1.1 Bacterial strains and sample processing

Three UPEC strains were isolated from the bladder biopsy specimens of three adult women with RUTIs. The patients were treated at the Urology Service of the Military Hospital for Specialties for Women and Neonatology from January 2017 to June 2021. Each patient signed an informed consent letter for the bladder biopsy procedure and urine sample collection.

The bladder biopsy specimens were obtained by the cystoscopy procedure, which was performed by a urologist. They were collected in sterile saline solution for biopsy culture, and collected in formaldehyde for pathological analysis. Bladder biopsy culture was performed at the Intestinal Bacteriology Research Laboratory from the Children's Hospital of Mexico "Federico Gómez." The biopsy tissue was mechanically ground in saline solution under sterile conditions and seeded in Petri dishes containing MacConkey agar (BD-BIOXON; Difco BD, 1 Becton Drive, Franklin Lakes, NJ. United States) and 5% sheep blood agar (BD-Difco Becton, Dickinson, and Company, Becton Drive Franklin Lakes, NJ, United States; De Nisco et al., 2019). Characteristic *E. coli* colonies were identified with the MALDI-TOF VITEK MS Microbial Identification System (bioMérieux, 376 Chemin de l'Orme, 69,280 Marcy-l'Étoile, France) at the HIMFG Central Clinical Laboratory. In the Department of Pathology of the National Institute of Neurology and Neurosurgery "Manuel Velasco Suárez," the pathological analysis of the biopsy specimens was carried out. The tissues were embedded in paraffin, cut with a microtome at 0.5 microns, and stained with hematoxylin-eosin (HE).

Urine samples were collected using the midstream technique and processed for culture in the Central Clinical Laboratory of HIMFG. The samples were seeded in Petri dishes containing cystine-lactose-electrolyte-deficient agar (CLED; BD-BIOXON), MacConkey (BD-BIOXON), and 5% sheep blood agar (BD-Difco) for the identification of classic uropathogens. The microbial count was performed and evaluated according to the Kass-Sandford criterion (Kass, 1956). *E. coli* strains were phenotypically identified using the MALDI-TOF VITEK MS Microbial Identification System (BioMérieux). Subsequently, the strains were propagated on trypticase soy agar plates (TSA; BD-Difco) and preserved in cryovials with trypticase soy broth (TSB; BD-Difco) supplemented with bovine fetal serum (BFS; ATCC, University Boulevard Manassas. VA, United States) at 1% and glycerol (Sigma–Aldrich, Spruce St. Louis, MO, United States) at 20%. The strains were kept at −70°C until use.

2.1.2 Pulsed-field gel electrophoresis assays of UPEC strains

The UPEC strains recovered from each patient's urine and bladder biopsy were evaluated by Pulsed-field gel electrophoresis (PFGE) to verify their clonal relationship. The PFGE assays were performed according to the protocols established by our working group (Ochoa et al., 2016; Contreras-Alvarado et al., 2021). The UPEC strains were embedded in low-melting-point (LMP) agarose blocks (Promega Corporation, Woods Hollow Road, Madison, WI, United States). The bacterial samples were digested using 20 U of the restriction enzyme Xba1 (New England Biolabs, Ipswich, MA, United States) at 37°C for 20 h. Subsequently, the PFGE was adjusted to run for 24 h at 200 V (7 V/cm) at an angle of 120°C to 14°C, with an initial pulse of 2.16 s and a final pulse of 13.58 s, in a CHEF Mapper system (Bio-Rad Life Science Research, Hercules, CA, United States). The macrorestriction patterns into the agarose gels were stained with GelRed® Nucleic Acid Stain (Biotium, Fremont, CA, United States), visualized with UV light and digitized using a CCD Camera Documenting System BK04S-3 (Biobase, Mexico City, Mexico). PFGE pulsotypes were analyzed with NTSYS pc v2.02j software, and the degree of clonality was evaluated considering the criteria described by Tenover et al. (1995). A lambda ladder PFGE marker (New England Biolabs, Hertfordshire, England, United Kingdom) was used as a molecular weight marker in this assay.

2.1.3 Haemolysis assays

The strains were seeded in Luria–Bertani (LB; DF-Difco) medium and cultured at 37°C overnight under constant agitation at 150 rpm. The suspension was then adjusted to 1.0 OD$_{600nm}$. To assess the presence of haemolysis, the bacteria were plated on agar supplemented with 5% sheep blood and incubated at 37°C for 18 h (Buxton, 2005).

2.1.4 Congo red uptake of UPEC strains

The strains were seeded in LB medium at 37°C overnight under constant agitation at 150 rpm. The suspension was adjusted to 1.0 OD$_{600nm}$ in fresh LB medium. Subsequently, 10 μL was dropped on YESCA agar plates (0.5 g/L yeast extract, 10.0 g/L casaaminoacids, 15.0 g/L bacteriological agar), and 1 mL of 1% Congo red (0.25 g) was added and incubated at 37°C for 72 h to evaluate the uptake of Congo red, an indicator of curli production (Reichhardt et al., 2015).

2.1.5 Motility tests

The strains were grown on LB agar plates (BD-Difco) at 37°C for 24 h. A colony of bacteria was incubated in tryptic soy broth (TSB) medium (BD-Difco) adjusted to an $OD_{600\ nm}$ value of 1.0 and then added to a tube with mobility-indole-ornithine medium (MIO; BD-Difco). The positive mobility in the tube was evaluated by the turbidity observed in the medium; negative motility bacteria were limited to the inoculation zone. The confirmatory mobility test was performed in 150 mm x 20 mm diameter Petri dishes containing semisolid mobility medium [0.25% agar (MCD LAB), 1% tryptone (Fluka Analytical), and 0.5% NaCl (J.T. Baker) and 0.005% 2,3,5- triphenyl tetrazolium chloride (TTC; MERCK)]. The medium was incubated with a bacterial colony from the LB plate at 37°C for 18 h. *E. coli* strain W3110 not motile (without halo in mm per 18 h of incubation), UPEC CFT073 motile (halos 24 mm/ 18 h) (Yang et al., 2016), and *Pseudomonas aeruginosa* ATCC 27853 extremely motile (halo 40 mm/ 18 h) (Saeki et al., 2021), were used as controls. The experiments were performed independently in triplicate. The motility obtained by the control strain CFT073 was used as a cut-off value to evaluate the motility of other strains. The UPEC strains were considered mobile when they presented motility halos ≥10 mm and immobile with a halo ≤9 mm.

2.1.6 Quantitative biofilm assays

UPEC strains were grown in TSB medium (BD-Difco) and incubated at 37°C for 24 h under constant agitation at 200 rpm. The 24-well microplates were prepared with 800 μL of TSB medium and 200 μL of the bacterial culture adjusted to an $OD_{600\ nm}$ of 1.0. After incubation of the microplate at 37°C for 24 h, three washes were performed with 1 mL of 1x PBS, and the biofilm was fixed with 1 mL of 4% formalin overnight at 4°C. The formalin was gently removed, and the biofilms were dried at room temperature and stained with 1% (w/v) crystal violet for 30 min. The excess dye was removed, the biofilms were washed with PBS 1x three times, and the microplate was allowed to air dry (Gomez et al., 2013). The dye was recovered using 1,000 μL of absolute methanol for 15–20 min in two stages, and the crystal violet retained by the biofilm was spectrophotometrically quantified at 620 nm in a microplate reader (MultiskanTM FC; Thermo Fisher, 81 Wyman Street, Waltham, MA 02451 United States). *E. coli* ATCC 25922, CFT073, and *E. coli* W3110 (lows biofilm former), *E. coli* J96 (moderate biofilm former), and *P. aeruginosa* ATCC 27853 (high biofilm former) (Stepanović et al., 2007; Gomez et al., 2013; Flores-Oropeza, 2017) were used as controls. The profiles of the biofilm-developing strains were categorized as high biofilm formers with ≥0.0180 OD_{620nm} values, moderate biofilm-forming strains with 0.0092–0.0179 OD_{620nm} values, low biofilm-forming strains with values 0.0053–0.0091 OD_{620nm}, and non-biofilm-forming strains with values ≤0.0052 OD_{620nm}. The experiments were performed independently in triplicate, and the results of all the experiments were plotted and statistically analyzed using GraphPad Prism version 8.0.0 for Windows (GraphPad Software, San Diego, California United States).[1]

2.1.7 HTB-5 cell adherence and invasion assays

The HTB-5 transitional cell carcinoma (TCC) cell line (ATCC, University Boulevard Manassas, VA, United States), which was isolated from the urinary bladder of a 67-year-old female with grade IV TCC, was cultured in culture flasks (Corning) with Eagle Minimum Essential Medium (EMEM; ATCC) and 10% bovine fetal serum (BFS; ATCC) up to 80% confluence. Previously, UPEC strains were grown in 5 mL of TSB medium (BD-DIFCO) to 37°C throughout the night with constant agitation (200 rpm). Next, 1×10^5 HTB-5 cells were seeded in a 24-well microplate in 1 mL of EMEM supplemented with 10% BFS/well. Prior to bacterial infection, the EMEM medium was added to the wells. The cell monolayers were infected at a multiplicity of infection (MOI) of 1:100, and 10 μL of a bacterial suspension adjusted to 1×10^7 bacteria/mL was added and incubated at 37°C for 3 h and 5% CO_2. Subsequently, bacteria not attached to the cell monolayers were removed with three gentle 1x PBS washes. For quantitative analysis, bacteria adhered to cell monolayers were collected after 1 min incubation with 1 mL PBS/0.1% Triton X-100, and serial dilutions (10^{-4}, 10^{-5}, and 10^{-6}) were grown in Petri plates with LB medium (BD-DIFCO) by the suspended drop method (10 μL) for the counting of CFUs. Finally, for counting invasive bacteria, 1 mL of EMEM medium was added to each well with 100 μg/mL gentamicin and incubated for 1 h. The wells were washed three times with 1x PBS to eliminate dead extracellular bacteria, 0.1% PBS/Triton X-100 solution was added, and the CFUs were counted in direct sample and three dilutions: 10^{-1}, 10^{-2}, and 10^{-3}. Cellular tests were performed independently in triplicate, and CFU/mL values were expressed as the average of the results from the three adherence assays. The averaged adherence and invasion data, along with the standard deviations, were graphed with GraphPad Prism version 8.0.0 for Windows. From the cut-off values of the UPEC strain CFT073, the adherence profiles of the UPEC strains were determined, such as highly adherent (≥2.72×10^6 CFU/mL), moderately adherent (1.18×10^6 – <2.72×10^6 CFU/mL), lowly adherent (8.15×10^5 – <1.18×10^6 CFU/mL), and nonadherent (< 8.15×10^5 CFU/mL). Following the same procedure, the invasive capacity of the UPEC strains was considered highly invasive (≥12.6 CFU/mL), moderately invasive (4.70 – <12.6 CFU/mL), low invasive (1.48 – <4.70 CFU/mL), and non-invasive (<1.48 CFU/mL).

2.1.8 Antibiotic susceptibility assays of UPEC strains

The minimum inhibitory concentration (MIC) was determined by the automated method of the VITEK 2 Systems based on the Therapeutic Microbiological Interpretation Guide. The Clinical and Laboratory Standards Institute (CLSI, 2022) guidelines (section 2.2) and Natural Resistance 2021 were used to determine the antibiotic susceptibility profile using 15 antibiotics belonging to 8 antibiotic categories: β-Lactam combination agents [Ampicillin-sulbactam (AMS), and Piperacillin-tazobactam (TZP)], Cephems 2nd generation: [Cefoxitin (FOX); Cephems 3rd generation: Ceftazidime (CAZ), and Ceftriaxone (CRO)], Cephems 4th generation: [Cefepime (CEF)], Carbapenems: [Doripenem (DOR), Ertapenem (ERT), Meropenem (MEM), and Imipenem (IPM)], Aminoglycosides: [Gentamicin (GM), and Amikacin (AMK)], Fluoroquinolones: [Ciprofloxacin (CIP)], Lipopeptides: [Colistin (COL)], and Glycylcycline: [Tigecycline (TIG)]. The reference strains *E. coli* ATCC 25922 (β-lactamase negative) and *E. coli* ATCC 35218 (ESBL-TEM-1 producer) were used as controls. The methodology for antibiotic

[1] www.graphpad.com

susceptibility was performed according to the guidelines of the CLSI (2022) and Ochoa et al. (2016). MDR strains are characterized by having acquired no susceptibility to at least one antibiotic in three or more classes. XDR strains have nonsusceptibility to at least one agent in all performed but two or fewer antibiotic classes (Magiorakos et al., 2012).

2.1.9 Phenotypic determination of ESBL and metallo-beta-lactamase in UPEC strains

Antibiotic disk susceptibility assays with CAZ (30 μg), aztreonam (ATM-30 μg), and CRO (30 μg) were performed for the presumptive identification of ESBL. In addition, antibiotic disks with MEM (10 μg) and IPM (10 μg) were used for metallo-beta-lactamase (MBL) identification, as suggested by CLSI (2022). Both enzyme groups were subjected to synergism assays with double and single disk assays. *Klebsiella pneumoniae* ATCC 700603 (ESBL+), *E. coli* ATCC 25922 (ESBL-), *P. aeruginosa* ATCC 27853 (MBL-), and clinical *P. aeruginosa* 540UC1 (MBL+) were used as control strains (Ochoa et al., 2015, 2016). Also, the phenotypic determination of ESBL production was performed using the VITEK 2 Systems automated method (Therapeutic Microbiological Interpretation Guide, CLSI section 2.2, and Natural Resistance 2021). The phenotypic assays were performed in triplicate.

2.2 Dry lab section

2.2.1 Sequencing and assembly of UPEC strains

The UPEC strains recovered from the bladder biopsy specimens were sent to the Sequencing Service of the Center for Genomic Sciences-UNAM. The genome was sequenced with the Illumina HiSeq platform (Illumina, Inc. 5,200 Illumina Way, San Diego, CA 92122, United States) and the Oxford Nanopore MiniION platform (Gosling Building, Edmund Halley Road, Oxford Science Park OX4 4DQ, United Kingdom). Hybrid assembly was performed *de novo* with Unicicler v. v0.4.1.[2] The assemblies were deposited in the BioProject database under the accession number PRJNA610084.[3]

2.2.2 Annotation and comparison of complete UPEC strain genomes.

The three assembled genomes were annotated with the NCBI Prokaryotic Genome annotation server[4] using the best-placed reference protein set method. In addition, they were annotated with GeneMarkS-2 + v4.11 (Thibaud-Nissen et al., 2016).[5]

The comparison of the complete genomes was performed using the MAUVE bioinformatics tools LASTZ and Geneious, included in the Geneious Prime 2023.0.3 software.[6] The loci with nucleotide differences, based on the alignments, were verified with SnapGene®-2023 software.[7] The hypothetical proteins were analyzed with InterProScan v5.65–97.0 software[8] (Paysan-Lafosse et al., 2023) to classify them by families and predict their functional domains. The infrequent regions identified during the WGS alignment and related to adhesion were called regions of interest.

2.2.3 Construction of the UPEC strain phylogenetic tree

In the NCBI[9] and BV-BRC[10] databases, a search for the genomes of "*E. coli*," from "Human" OR "*Homo sapiens*," was performed with the following characteristics: sequence status "Complete," source of isolation "urine" OR "urinary tract," and disease "urinary tract infection" OR "UTI" OR "cystitis" OR "pyelonephritis." The final database was manually curated. In addition, 160 genomes associated with the urinary tract were selected, and 11 genomes were characterized as controls [*E. fergusonii* ATCC 35469 (CU928158), EAEC/STEC 042 (FN554766), EHEC Sakai (CP028307), EIEC CFSAN029787 (CP011416), EPEC XH987 (CP102675), ETEC H10407 (FN649414), K12 MG1655 (U00096), K12 W3110 (CP017979), NMEC S88 (CU928161), *Shigella flexneri* 2a ATCC 29903 (CP026788), and STEC 00–3,076 (CP027584)], through the service of BV-BRC 3.30.19 Bacterial Genome Tree.[11] For the construction of the phylogenetic tree, the file in Newick format (.nwk) was exported to iTOL v6.8,[12] for annotation, and finally, the Scalable Vector Graphics file (.svg) was edited in the Inkscape v1.2.2 program.[13]

2.2.4 Identification of virulence, fitness and resistance genes

The search for virulence and resistance genes (antibiotics, heavy metals, and quaternary salts) in chromosomes and plasmids was performed using the VirulenceFinder v2.0 servers,[14] VRprofile2 v2.0,[15] OriTFinder 1.1,[16] VICTORs and PATRIC from the Bacterial and Viral Bioinformatics Resource Center (BV-BCR) 3.28. 9,[17] ResFinder 4.1[18] and The Comprehensive Antibiotic Resistance Database.[19] The results were compared using the CLUSTAL Omega bioinformatics tools v1.2.22 in the Geneious Prime software, UNIPROT 2023[20] (The UniProt Consortium, 2023) and NCBI.[21]

2.2.5 Identification of phylogenetic groups, genomic Islands and prophages

The phylogenetic group of the UPEC strains was identified using the EzClermont 0.7.0 command tool using the following link: https://ezclermont.hutton.ac.uk/ (Beghain et al., 2018; Clermont et al., 2019). The search for Genomic Islands as elements capable of transferring

[2] https://github.com/rrwick/Unicycler
[3] https://www.ncbi.nlm.nih.gov/bioproject/610084
[4] https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
[5] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/
[6] https://www.geneious.com
[7] from Dotmatics; available at snapgene.com
[8] https://www.ebi.ac.uk/interpro/search/sequence/
[9] https://www.ncbi.nlm.nih.gov/
[10] https://www.bv-brc.org/
[11] https://www.bv-brc.org/app/PhylogeneticTree
[12] https://itol.embl.de/
[13] https://inkscape.org/
[14] https://cge.food.dtu.dk/services/Virulen
[15] https://tool2-mml.sjtu.edu.cn/VRprofile/VRprofile.php
[16] https://bioinfo-mml.sjtu.edu.cn/oriTfinder/
[17] https://www.bv-brc.org/
[18] https://cge.cbs.dtu.dk/services/ResFinder/
[19] CARD, https://card.mcmaster.ca/home
[20] https://www.uniprot.org/
[21] https://www.ncbi.nlm.nih.gov/

TABLE 1 Clinical description of patients with RUTIs obtained of the military hospital for specialties for women and neonatology in Mexico City, Mexico.

Patient	P1	P2	P3
Strain	UTI-1_774U	UTI-2_245U	UTI-3_455U
Age (sex)	50	30	60
Clinical antecedents	Bullous cystitis	Interstitial cystitis, diabetes, lumbar trauma, and voiding dysfunction	Cyclic neutropenia and urinary incontinence.
Initial treatment date	2017–01	2018–03	2018–12
Disease evolution (years)	3.25	2.0	1.58
Last antibiotic treatment	Amoxicillin/clavulanic acid	ND	Phenazopyridine/nitrofurantoin
Systemic treatment	Intravesical installation of hyaluronate (12 doses)	Intravesical installation of hyaluronate (12 doses)	ND
Prophylaxis, replacement, and psychological treatment	Oral lactobacilli and nitrofurantoin	Anticholinergic and behavior therapy	Oral lactobacilli, topic estrogens, and behavior therapy
Biological sample	Concentrated urine in 1000 CFU/mL and bladder biopsy	Urine culture 100,000 CFU/mL, bladder biopsy	Urine culture 100,000 CFU/mL and bladder biopsy

P1 (Patient 1); P2 (Patient 2); P3 (Patient 3); UTI, Urinary Tract Infection; UPEC, Uropathogenic Escherichia coli; CFU/mL (Colony Forming Unit per milliliter); ND, nondate.

virulence between strains was carried out using IslandViewer v.4.0[22] (Bertelli et al., 2017) and VRprofile v.2.0 (see footnote 15, respectively; Wang et al., 2022). The presence of prophages was determined by employing the PHASTER 4.3X platform[23] (Zhou et al., 2011; Arndt et al., 2016).

2.2.6 Determination of UPEC serotypes, sequence types, and clonal complexes

The serotypes of the UPEC strains were evaluated with the SeroTypeFinder 2.0 server[24] for the identification of the following antigens: somatic (O) and flagellar (H) (Joensen et al., 2015). The STs of the UPEC strains were determined by allele identification in the sequences of seven "Housekeeping" genes suggested by the University of Oxford's Multilocus Sequence Type (MLST) web platform. The alleles in the gene sequences *adk* (adenylate kinase), *fumC* (fumarate hydratase), *gyrB* (DNA gyrase), *icd* (isocitrate/isopropyl malate dehydrogenase), *mdh* (malate dehydrogenase), *purA* (adenylosuccinate dehydrogenase), and *recA* (ATP/GTP binding motif) were identified with the platform[25] (Jolley et al., 2018).

3 Results

3.1 Phenotypic characterization

3.1.1 Patient clinical-pathological data

UPEC strains were recovered from 3 adult women with RUTIs for at least 1.5 years treated at the Urology Service of the Military Hospital for Specialties for Women and Neonatology, Secretary of National Defense, Mexico City. Bladder biopsy specimens were collected from each patient 1 month after concluding antimicrobial and topical treatment. Patient 1 (P1) was diagnosed with follicular or bullous cystitis and had no history of comorbidities. Patient 2 (P2) presented a diagnosis of interstitial cystitis and suffered a lumbar trauma that led to a bladder emptying disorder; patient 2 also presented type 2 diabetes mellitus. Patient 3 (P3) was diagnosed with moderate stress mixed urinary incontinence and cyclic neutropenia according to clinical data. The clinical strains of RUTIs recovered from the patients were called UTI-1_774U (P1), UTI-2_245U (P2), and UTI-3_455U (P3). The patients received treatment with multiple broad-spectrum antibiotics, such as amoxicillin/clavulanic acid and nitrofurantoin, over the course of infection. In addition, nonantibiotic topical treatment was applied with intravesical instillations of hyaluronate, anticholinergics, and local oestrogens. The clinical data of each patient, including age, antibiotic/nonantibiotic treatment, topical application, recovered strain, and biological evaluation of the clinical sample, are shown in Table 1 and Supplementary Figure S1.

The histopathological micrographs of the three-patient bladder biopsy specimens showed a severe inflammatory response (Figure 1). The bladder tissue of P1 showed an intense lymphocytic inflammatory infiltrate in the submucosa with discrete exocytosis of lymphocytes through the urothelial epithelium (Figures 1A,B). The bladder tissue of P2 also showed severe chronic lymphocytic inflammation with total involvement of the mucosa, including mucosal detachment (Figure 1C). Significant inflammatory cellularity was observed in the submucosa (Figure 1D). In the bladder tissue of P3, chronic nonspecific lymphocytic inflammation in the submucosa was demonstrated, and focal detachment of the epithelium was also observed (Figures 1E,F).

3.1.2 Haemolytic activity, adherence, and invasion of UPEC strains

The haemolytic activity of the UPEC strains UTI-1_774U, UTI-2_245U, and UTI-3_455U visualized in agar supplemented with 5–10% sheep blood showed a haemolytic activity similar to that of the negative control *E. coli* strain W3110 (Supplementary Figure S2). Moreover, the genes associated with haemolytic activity in the tested strains are listed in Supplementary Table S1. The UPEC strain CFT073 used as a positive control showed haemolytic activity in a β-haemolysis zone around the inoculation site (Supplementary Figure S2).

[22] https://www.pathogenomics.sfu.ca/islandviewer/
[23] https://phaster.ca/
[24] https://www.ridom.de/u/E._coli_CGE_SerotypeFinder.html
[25] https://pubmlst.org/escherichia/; accessed on 13 October 2021.

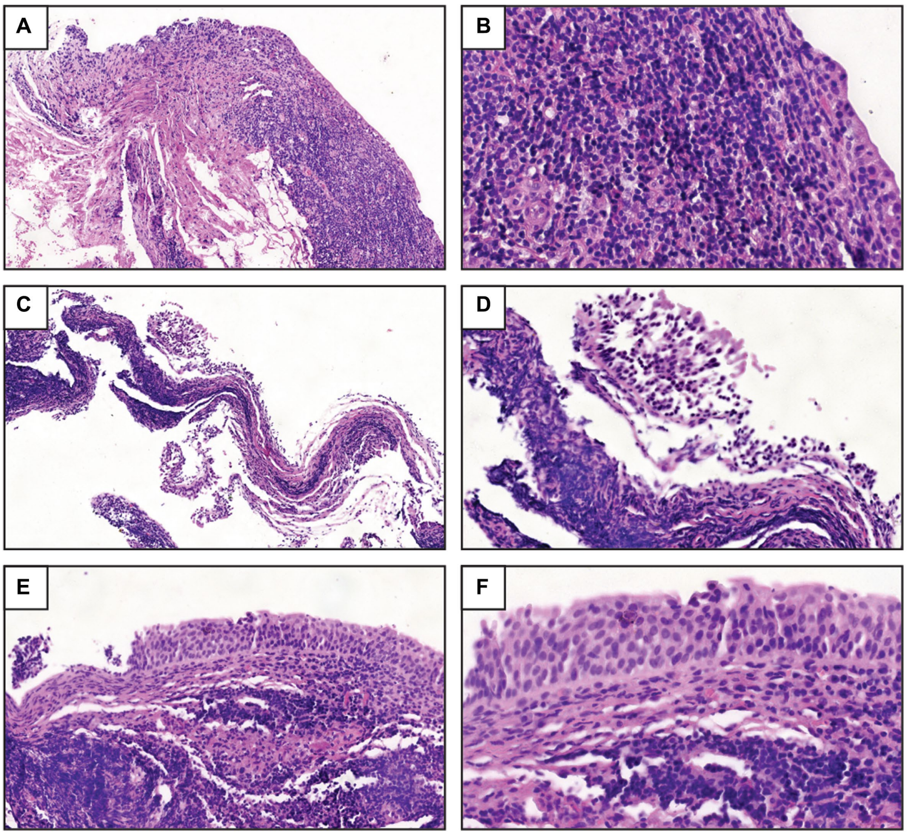


FIGURE 1
Histopathological study of the biopsy samples. **(A)** UTI-1_774U biopsy sample obtained from women with RUTIs showed a severe inflammatory response and lymphocytic infiltration in the submucosa [hematoxylin-eosin stain (HES), 10X]. **(B)** UTI-1_774U biopsy sample with discrete exocytosis of lymphocytes through the urothelial epithelium (HES, 40X). **(C)** UTI-2_245U biopsy sample with intense chronic inflammation and mucosal detachment (HES, 10X). **(D)** Marked cellularity and inflammatory response in the submucosa (HES, 20X) of UTI-2_245U biopsy sections. **(E)** UTI-3_455U biopsy sample with chronic nonspecific lymphocytic inflammation in the submucosa (HES, 20X). **(F)** UTI-3_455U biopsy sample with focal epithelium detachment (HES, 40X).

Quantitative analysis of adherence showed that the UTI-1_774U strain (3.15×10^6 CFU/mL) was significantly ($p<0.0001$) highly adherent to HTB-5 cells in comparison with the strains UTI-3_455U (1.85×10^6 CFU/mL), UTI-2_245U (1.34×10^6 CFU/mL), UPEC CFT073 (1.54×10^6 CFU/mL), and *E. coli* W3110 (2.36×10^6 CFU/mL), as shown in Figure 2A. The quantitative analysis of invasion showed that the UTI-3_455U strain (11 CFU/mL) was significantly invasive to HTB-5 cells when compared to the strains UTI-1_774U (8.72 CFU/mL), UTI-2_245U (9.33 CFU/mL), UPEC CFT073 (7.92 CFU/mL), and *E. coli* W3110 (9.15 CFU/mL). In addition, complete genome analysis using the PATRIC and NCBI platforms showed the presence of genes associated with adherence and invasion in the strains (Supplementary Figure S1).

The presence of fimbriae potentially involved in the adhesion and invasion processes under the described conditions was subsequently evaluated by TEM. Negative staining TEM micrographs with 0.75% uranyl acetate showed the presence of short (>1 µm) and thin fimbriae in UTI-2_245U. However, in the micrographs of strains UTI-1_774U and UTI-3_455U, the fimbria were not observed under these growth conditions. The control strains of *E. coli* W3110 and UPEC CFT073 also produced shorter and thinner fimbriae than those of UTI-2_245U (Figure 2). On the other hand, the three recurrent UPEC strains assembled the curli fimbriae when cultured overnight, at 37°C and 20°C, and YESCA medium (data not shown).

3.1.3 Biofilm formation by UPEC strains

The formation of biofilms by the spectrophotometric quantification method of the crystal violet dye retained by the biofilm at 620 nm showed that the UPEC strains UTI-1_774U ($OD_{620\,nm}=0.0003$), UTI-2_245U ($OD_{620\,nm}=0.0019$) and UTI -3_455U ($OD_{620\,nm}=0.0018$) were not biofilm formers when compared with the positive control strains *Pseudomonas aeruginosa* ($OD_{620\,nm}=0.0981$) and UPEC J96 ($OD_{620\,nm}=0.0143$) and with the negative control strains UPEC CFT073 ($OD_{620\,nm}=0.0034$) and *E. coli* ATCC 25922 ($OD_{620\,nm}=0.0043$) (Supplementary Figure S3). The bioinformatic analysis of the sequenced genomes showed the presence of genes associated with biofilm formation in the UPEC strains. These genes are shown in Supplementary Table S1.

3.2 Genotypic characterization

3.2.1 General characteristics of the UPEC genomes

The sequence analysis of the three UPEC strains showed that the genomes were different sizes: 5.19 Mb for UTI-1_774U, 5.01 Mb for UTI-2_245U, and 5.0 Mb for UTI-3_455U. Comparative analysis of the three UPEC genomes using the local pairwise alignment approach showed 99% overall identity between UTI-1_774U and UTI-3_455U,

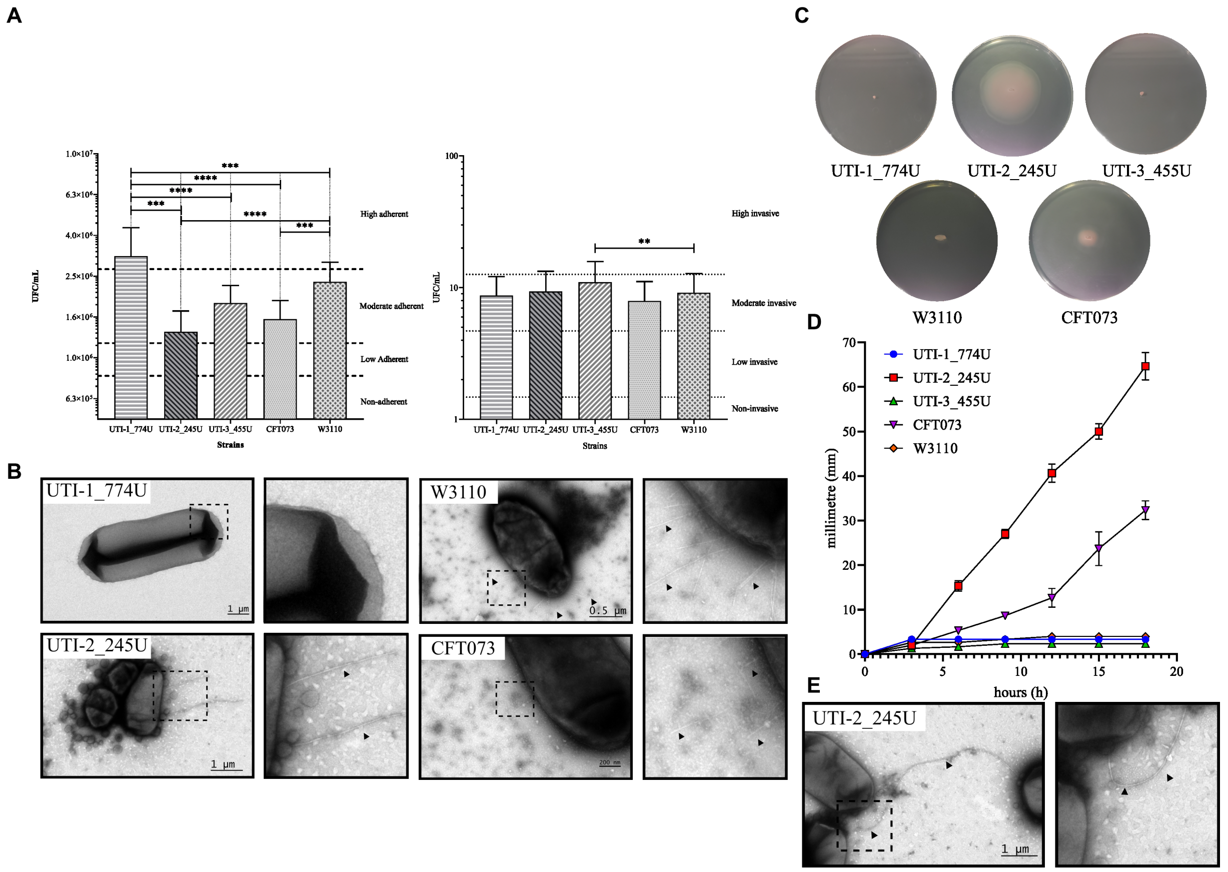


FIGURE 2
Phenotype assays of RUTI strains. **(A)** Adherence and invasion assays of HTB-5 human bladder cells. RUTI strains are adherent and invasive; the UTI-1_774U strain is highly adherent, while the UTI-3_455U strain is moderate invasive. **(B)** The fimbria of the RUTI strains and the controls are visualized by TEM. **(C)** Mobility assays and mobility plates with tetrazolium chloride show a halo of mobility around the colony. UPEC strains UTI-1_774U and UTI-3_455U are not motile. The mobility speed graph shows that the UTI-2_245U strain is highly mobile. **(D)** Quantitative analysis of the motility by flagella of the three clinical strains. UTI-2_245U strain shown a motility developed. **(E)** TEM micrography showing the flagella in the UTI-2_245U strain.

93% identity between UTI-2_245U and UTI-3_455U, and 88% identity between UTI-1_774U and UTI-2_245U (Supplementary Table S2). Comparative analysis of the whole genomes showed the lack of PAIs in the UTI-1_774U, UTI-2_245U, and UTI-3_455U strains when compared with UPEC CFT073 (Supplementary Figure S4). The whole genome of the UTI-1_774U strain contained 4,954 genes, 160 pseudogenes, and 292 genes that code for hypothetical proteins (Table 2). This strain harbors a cryptic plasmid of 2,113 bp, a nonconjugative mobilization plasmid of 4,072 bp, a conjugative mobilization plasmid of 5,631 bp, and a resistance carrier plasmid of 102,591 bp (Table 2). All data obtained from the bioinformatics databases of the 172 genomes included in this study are displayed in the Excel Supplementary Database.

The genome of the UTI-2_245U strain contains 5,205 genes, 153 pseudogenes, and 280 genes encoding hypothetical proteins. In addition, it contains six plasmids, a 5,119 bp colistin production-associated plasmid, a 5,631 bp conjugative mobilization plasmid, a 6,336 bp resistance plasmid, a 68,488 bp fertility plasmid, a phage plasmid from 90,408 bp, and a resistance plasmid from 109,478 bp (Table 2; Excel Supplementary Database). The genome of the UTI-3_455U strain contains 5,049 genes, 160 pseudogenes, and 296 genes encoding hypothetical proteins. Furthermore, it contains five plasmids, two small cryptic plasmids (1,822 bp and 2,113 bp), a 5,630 bp conjugative mobilization plasmid, an 88,163 bp conjugation plasmid, and a resistance plasmid of 91,528 bp (Table 2; Excel Supplementary Database). The plasmids found in the UPEC strains associated with RUTIs, UTI-1_774U, UTI-2_245U, and UTI-3_455U, are shown in the Supplementary Figure S5. The UTI-2_245U genome contains the *cas* I-E operon and two CRISPR arrays with 13 spacers, similar to the genomes of several enterobacteria, mainly the serotypes of other *E. coli* and serogroups of *Shigella* spp. Interestingly, high similarity to the bacteriophage *Myoviridae* sp. ctitt1 was also identified. In contrast, the UTI-1_774U and UTI-3_455U strains did not display CRISPR *Cas* genes; however, four repeat and spacer regions with high sequence similarity to *Enterobacteriaceae* were identified (Supplementary Table S3).

3.2.2 Phylogenetic comparison of UPEC strains

The comparative phylogenetic analysis included 157 strains of *E. coli* associated with urinary infection, 11 strains of *E. coli* control strains, and the three isolated RUTIs strains. These strains were distributed into nine well-defined clusters according to their phylogenetic group, STs, serotype, and their origin of isolation (Figure 3). In cluster 1, 33 clinical strains of urinary *E. coli* and three control strains (ETEC H10407, *E. coli* K12 W3110, and *E. coli* K12 MG1655) were distributed. In addition, 52.77% (19/36) of them were

TABLE 2 General description of the genomes of the three UPEC strains associated to RUTIs.

Strain	Source of DNA	GenBank®	Size (bp)	Genes (Total)	CDSs (Total)	Genes (coding)	Pseudogenes	Hypothetic proteins	Plasmid function
UTI-1_774 U	Chromosome	CP049852.1	5,192,928	4,954	4,835	4,675	160	292	
	pEcoUTI1a	CP049856.1	2,113	2	2	1	0	1	Small cryptic
	pEcoUTI1b	CP049855.1	4,072	3	3	3	0	0	Non-conjugative mobilization
	pEcoUTI1c	CP049854.1	5,631	7	7	4	2	1	Conjugative mobilization
	pEcoUTI1d	CP049853.1	102,591	117	117	46	56	15	Resistance
UTI-2_245 U	Chromosome	CP049845.1	5,016,451	5,205	5,080	4,927	153	280	
	pEcoUTI2a	CP049851.1	5,119	10	10	5	2	3	Colicin production
	pEcoUTI2b	CP049850.1	5,631	7	7	4	2	1	Conjugative mobilization
	pEcoUTI2c	CP049849.1	6,336	8	8	1	6	1	Resistance
	pEcoUTI2d	CP049848.1	68,488	89	89	67	2	20	Fertility
	pEcoUTI2e	CP049847.1	90,408	107	104	61	16	30	Extrachromosomal-phage P1
	pEcoUTI2f	CP049846.1	109,478	126	126	90	10	26	Resistance
UTI-3_455 U	Chromosome	CP049839.1	5,003,672	5,049	4,930	4,770	160	296	
	pEcoUTI3a	CP049843.1	1,822	2	2	2	0	0	Small cryptic
	pEcoUTI3b	CP049842.1	2,113	2	2	1	0	1	Small cryptic
	pEcoUTI3c	CP049841.1	5,630	7	7	6	0	1	Conjugative mobilization
	pEcoUTI3d	CP049844.1	88,163	97	97	78	2	17	Conjugative
	pEcoUTI3e	CP049840.1	91,528	103	103	38	48	17	Resistance

CDS, Coding Sequence; pEcoUTI, plasmid from *Escherichia coli* urinary tract infection strain.

isolates mainly of Asian descent, 100% (36/36) belonged to phylogenetic group A, 36.11% (13/36) belonged to ST167, 52.77% (19/36) belonged to serogroup O101, and 30.55% (11/36) contained the H10 flagellar antigen. The predominant serotypes were O101:H10, 30.55% (11/36), and O101:H9, 16.66% (6/36). In cluster 2, three UPEC strains were isolated from Asia, all belonging to phylogenetic group A. Thirteen strains were grouped in cluster 3, and 38.46% (5/13) were from Asia; while 23.07% (3/13) were from North America and Africa. Moreover, 100% (13/13) of these stains belonged to phylogenetic group C, 84.61% (11/13) belonged mainly to ST410, and 76.92% (10/13) had the H9 flagella antigen. The most frequent serotype was O8:H9 in 23.07% (3/13) (Figure 3). Five clinical strains of UPEC were grouped in cluster 4, and 60% (3/5) were of South American origin. Moreover, 80% (4/5) belonged to the phylogenetic group B1, ST224, and flagella antigen H23. The strain EC10 O157:H7 recovered from urine was included in this cluster. Twenty-eight clinical strains of UPEC and four control strains (EIEC CFSAN029787, EPEC XH987, STEC 00–3,076, and *E. coli* Sakai O157:H7) were grouped in cluster 5, 62.5% were of Asian origin and all belonged to phylogenetic group B1. In this cluster, *E. coli* strains from urinary infections with a wide variety of "O" (O23, O9, O160, O:82, O:188, O:154, and O157) and "H" (H25, H:16, H:31, H8, H9, and H21) antigens and ST (ST6388, ST453, ST101, and ST2179) were grouped. In addition, the RUTIs clinical strain UTI-2_245U was grouped, with serotypes O141:H5 and ST295 and phylogenetic group B1. In this clade, strains related to UTIs and urinary *E. coli* were grouped (Figure 3). In cluster 6, five clinical strains of UPEC and two control strains (*Shigella flexneri* 2a ATCC 29903 and *Escherichia fergusonii* ATCC 35469) were distributed. Additionally, 42.85% (3/7) belonged to phylogenetic groups E, ST219, ST2847, serotype O51:H34, and H28. In this cluster, a clinical strain of phylogroup A (O13:H16/ST245) and a clinical strain of phylogroup D (O86:H51/ST1011) were also grouped. In cluster 7, 13 clinical strains of UPEC and one control strain (EAEC_STEC_142) were grouped. In addition, 100% (14/14) belonged to phylogenetic group D, and 50% (7/14) belonged to serogroup O102. Of these strains, 57.14% (8/14) showed a flagellar antigen H6, 42.85% (6/14) were ST405, and 50% (7/14) were of North American origin (Figure 3). Eight clinical strains of UPEC were grouped in cluster 8, and 87.5% (7/8) belonged to phylogroup F, while 50% (4/8) were of North American and Asian origin. Of them, 62.5% (5/8) belonged to ST648, 50% (4/8) belonged to two serogroups (O1 and O153), and 37.5% (3/8) were flagellar antigen H6. In this cluster, one strain belonging to phylogenetic group G showed serogroup O24 and was of North America. Finally, in cluster 9, 52 UPEC strains and a control strain of

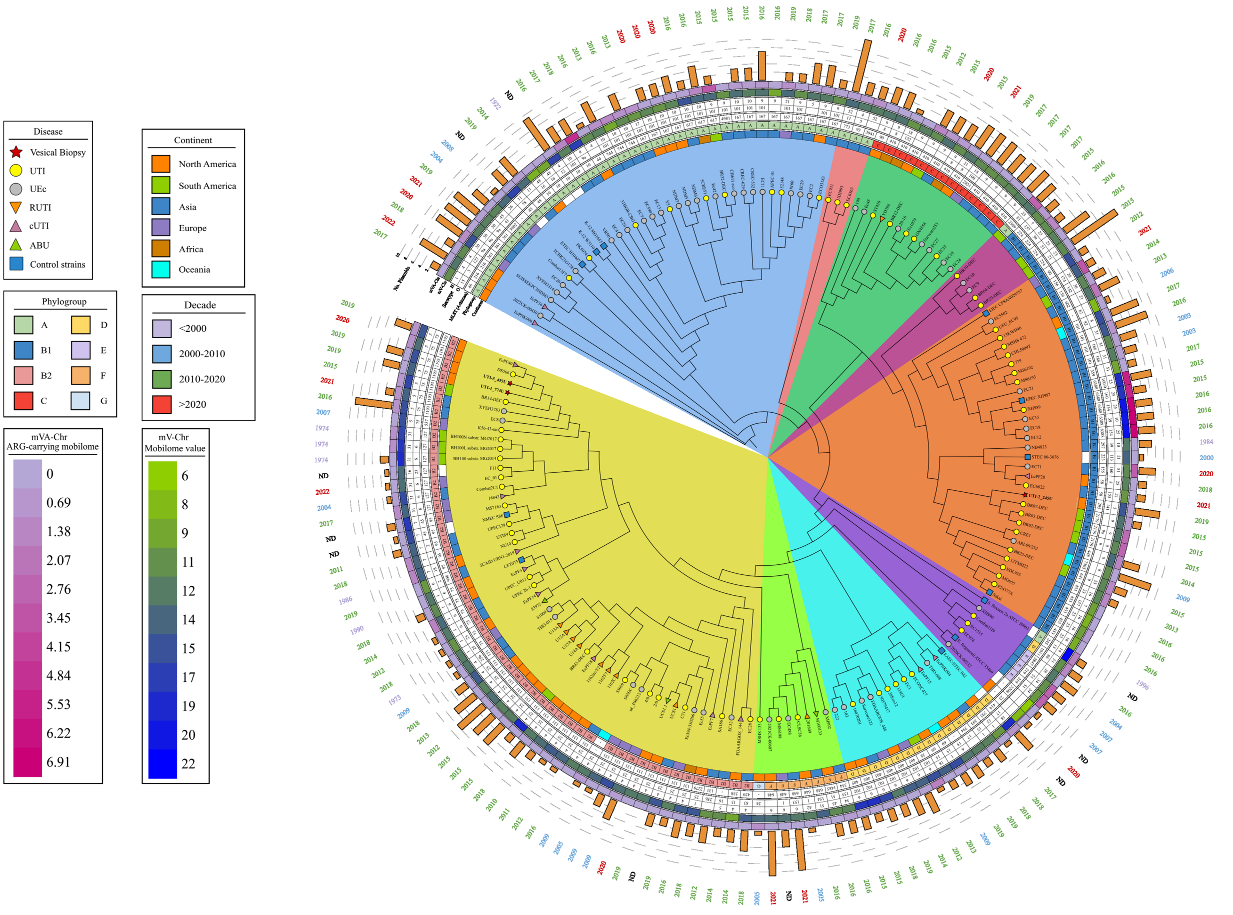


FIGURE 3
Phylogenetic map of 172 uEC strains and controls. The phylogenetic tree obtained through the BV-BCR 3.30.19 service, Bacterial Genome Tree, is presented. The disease data, phylogroup, continent, date of isolation, and mobilome of the strains are presented, and colored shadows represent the clades obtained. Cluster 1 (blue), cluster 2 (pink), cluster 3 (green), cluster 4 (violet), cluster 5 (orange), cluster 6 (purple), cluster 7 (light blue), cluster 8 (lime green), and cluster 9 (yellow).

extraintestinal *E. coli* (NMEC 588) were grouped. Of these strains, 100% (53/53) belonged to phylogenetic group B2, 43.39% (23/53) were associated with serogroup O25, 43.39% (23/53) showed the flagellar antigen H4, 39.62% (21/53) were ST131, 37.73% (20/53) were of North America, and 28.30% (15/53) were of Asian origin. The most frequent serotypes in this cluster were the following: strains 33.73% (20/53) were O25:H4/ST131, 13.20% (7/53) were O75:H5/ST1193, 9.42% (5/53) were O6:H31/ST127, and 7.54% (4/53) were O6:H1/ST73. The two RUTIs strains (UTI-1_774U and UTI-3_455U) were grouped in this cluster and belonged to phylogenetic group B2 and serotypes O75:H5 and ST1193. The strains in this cluster were associated with UTIs, uEC strains, and complicated UTIs (cUTI) and were mainly localized in the bladder (Figure 3). The correlation of the following frequencies: disease vs. STs, disease vs. region, and disease vs. number of plasmids, are shown through heat maps (Supplementary Figure S6).

3.2.3 Analysis of the UPEC strain STs

From the comparative genome analysis, the UTI-1_774 U and UTI-3_455 U strains strongly correlated with 24 *E. coli* strains from different origins. These strains belonged to ST1193, a widely-distributed clone that is a causative agent of human infections and healthcare-associated infections (HCAIs) in hospitals around the world. The ST1193 clone belongs to the B2 phylogenetic group and the O75:H5 serotype. Generally, this clone can be isolated from RUTIs cultures, cerebrospinal fluid (CSF), blood, catheter cultures, and it is also present in healthy individuals. The bioinformatic analysis showed a high proportion of genes associated with the mobilome in strain ST1193; however, the proportion of genes associated with resistance was low (Figure 4A).

The UTI-2245 U strain showed a strong association with six *E. coli* strains of different origins belonging to ST295 (Figure 2B). Comparative genome analysis indicated that this clone has an animal origin, mainly domestic and farm animals. The ST295 strains belong to two commensal phylogenetic groups (B1 and A), three serogroups (O10, O16, and O141), and the flagellar variant H16. Strains belonging to ST295 contained a low proportion of genes associated with the mobilome; in contrast, a high proportion of genes was associated with resistance in strains belonging to ST1193 (Figure 4B).

3.2.4 Comparison of complete UPEC strain genomes

The phylogenetic tree was built from the three clinical strains of RUTIs of bladder biopsy culture included in this study, 13 cUTI

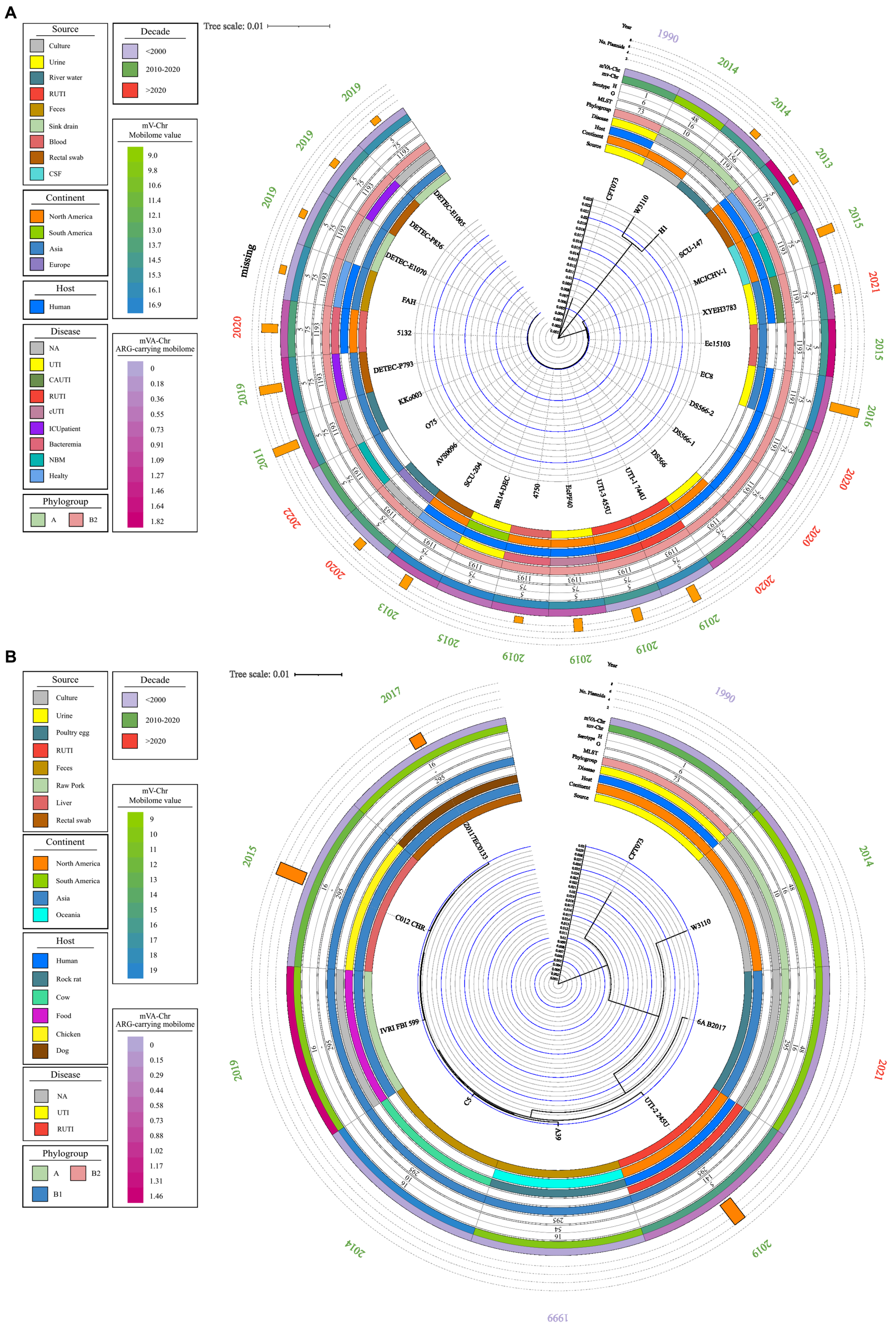


FIGURE 4
Phylogenetic map of ST1193 and ST295. A phylogenetic map was prepared using the complete genome using MAUVE and Geneious Prime. Data on source, disease, and continent of origin are presented, following the color code in the legend key. The year of isolation, phylogroup, serotype, ST, and the number of plasmids in each genome are also shown. **(A)** Phylogenetic tree of ST1193. **(B)** Phylogenetic tree of ST295.

genomes, 10 RUTI genomes, and 11 control genomes, all previously described in this study. The phylogenetic tree of the 36 genomes generated with MAUVE and Geneious identified *E. fergusonii* as a clonally unrelated strain (outlier). Moreover, 97.22% (35/36) of the remaining genomes showed 99.99 identity, and two main clades were generated. Clade A included the control strains of intestinal origin (EAEC_STEC_042, EIEC_ CFAN029787, EPEC_ XH987 and ETEC_ H10407), *S. flexneri* and two *E. coli* K12 strains (K12_MG1655 and *E. coli* W3110) (Figure 5). The genomes in clade A shared the following regions: operon *yhE* fimbria-W3110, operon *ybg* -W3110, operon class 5 fimbria-like of *cfaE*/*cblD* family-W3110, operon cryptic of the fimbria *yra*-W3110, operon class 5 fimbria-like of *cfaE*/*cblD* of UTI-2, fimbria *elf*/*ycvb* -W3110, operon *fim* fimbriae of CFT073, and adhesin and invasin-New PAI of UTI-1. Interestingly, the *csg*, *yeh*, *ecp*/*mat*, *fdeC intimin-like*, *fim*, *yfc*, *F9*/*yde*/*fml*, *ybg*, *sfm*, *gaf-F17-like*, *yqi*, *and cfa*/*cdI* operons were grouped together in the clade; however, in this clade, the *yad*, *auf*, *pap*, *sfa*/*foc*, and *afa*/*drA* operons were not identified. Clade A contains genomes mainly associated with phylogroups A and B1. The genomes associated with this clade showed a heterogeneous distribution of serotypes and STs (Figure 5).

Clade A was divided into four subclades, A1, A2a, A2b and A2c. Clade A1 was made up of three strains belonging to phylogenetic group D, and the intestinal strain EAEC_STEC_042 was included. Clade A2a contained *Shigella flexneri*. Clade A2b included the RUTIs UTI-2_245U and YD786, a cUTI EcPF20, EIEC_CFAN029787, and EPEC_XH987. The genomes of this clade mainly belonged to phylogenetic group B1 and were associated with the *ehaA* adhesin and the *cfa*/*cbi* fimbriae operon. In this clade, the RUTI strain UTI-2_245U had a high number of plasmids. Two cUTI genomes (Ec PF16 and EcPNK006) and three control strains (ETEC_ H10407, K12_MG1655, and *E. coli* W3110) were grouped into A2c. This clade did not have homologous regions of interest, and the genomes were characterized by the presence of intimin-like *fdeC,* which were absent in the control genomes. However, fimbrial operons were absent, denoted by *ecp*/*mat*, *gaf* F17-like (Figure 5).

Clade B was mainly characterized by the presence of nonhomologous regions of interest. In these regions, genes encoding the type 1 fimbria-lke of UTI-1, Adhesin-Invasin New PAI of UTI-1, *yqi* fimbria of CFT073, and regulatory integrase of type 1 fimbria *hxy* of UTI-1 were found. Clade B was made up of 21 urinary *E. coli* strains (8 cUTI strains, 11 RUTIs strains, one NMEC_S88 and UPEC CFT073), 95.23% (20/21) belonging to phylogenetic group B2 and 71.42% (15/21) showing three ST types [14.28% (3/21) for ST73, 42.85% (9/21) for ST131, and 14.28 (3/21) for ST1193]. Interestingly, the genome of RUTI 201609 categorized into phylogenomic group F was not included in clade B with the other RUTIs, genomes (Figure 5).

Clade B is divided into three subclades, B1, B2a, and B2b. Clade B1, containing only strain 201,609, is characterized by the presence of the *cfa*/*cbi* operons; however, there is an absence of the *ybg* and sfm fimbrial operons. Clade B2a includes the strains UTI-1_774U and UTI-3_455U, characterized by the element's fimbria type 1-like of UTI-1, *upaG* of CFT073, *ehaB* of UTI-2, sugar transporter, operon *ygb*

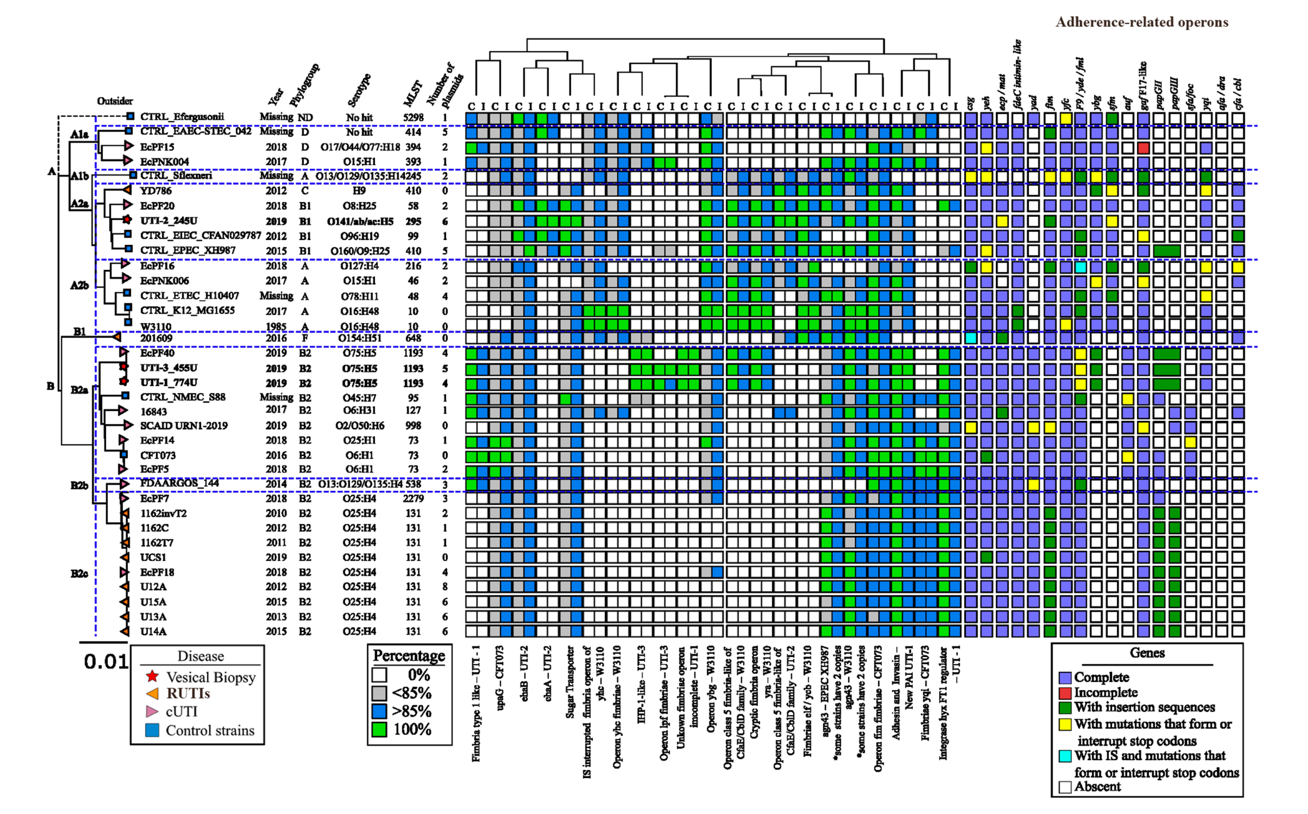


FIGURE 5
Identification of genes and operons associated with adherence in *E. coli* strains. The values obtained by BLAST of the regions of interest are presented. The percentage of identity or coverage is represented in colored boxes using the color code shown in the "Percentage" legend key. In the analysis of the operons associated with fimbriae, each box represents an operon, and the color following the key present in the legend "Operons" shows if it is complete, partial, or contains point mutations that generate premature stop codons or insertion sequences.

of W3110, Agn43 of EPEC_XH987, Adhesin-Invasin New PAI of UTI-1, and the regulatory integrase of the type 1 fimbria *hxy* of UTI-1. Additionally, the genomes of this clade had the fimbrial operons *csg*, *yeh*, *ecp/mat*, *fdeC intimin-like*, *yad, fim*, *yfc*, *F9/yde/fml*, *auf,* and *gaf-F17- like*. The strains in clade B2a were mainly from serotype O75:H5/ST1193 [33.33% (3/9)]. The B2b clade mainly included the recurrent strains and is characterized by having the same pattern of nonhomologous regions of interest and fimbrial patterns strains, but does not have the fimbria type 1-like of UTI-1, operon *ygb* of W3110 and *papG*II and *papG*III fimbrial operons. The strains in clade B2b were mainly from serotype O25:H4/ST131 [81.81% (9/11)] (Figure 5).

3.2.5 Phenotypic analysis of the antibiotic resistome in the three RUTI-associated UPEC strains

The susceptibility profile to 14 categories of antibiotics allowed us to determine the phenotypic antibiotic resistome in the UPEC strains associated with RUTIs. The UPEC strain UTI-1_774U displayed an MDR-6 profile (multidrug resistance to 6 antibiotic categories including penicillin, β-lactam combination agents, lipopeptides, macrolides, quinolones and fluoroquinolones, and folate pathway antagonists). The UPEC strain UTI-2_245U showed an MDR-5 profile (multidrug resistance to 5 antibiotic categories including β-lactam combination agents, lipopeptides, macrolides, tetracyclines, and folate pathway antagonists). The UPEC strain UTI-3_455U showed an MDR-3 profile (multidrug resistance to 3 antibiotic categories including lipopeptides, macrolides, quinolones, and fluoroquinolones). None of the three UPEC strains were phenotypic ESBL producers (Table 3).

The resistome analysis of the strains UTI-1_774U, UTI-2_245U, and UTI-3_455U presented 84, 86, and 80 genetic elements associated with antibiotic resistance, respectively. Additionally, 59.52% (50/84) for UTI-1, 61.62% (53/86) for UTI-2, and 63.75% (51/80) for UTI-3 display mobile elements associated with active efflux pumps. In the UTI-1_774U strain, unique genetic elements were found, such as aminoglycoside O-phosphotransferase (*APH(6)-Id*), aminoglycoside 3′-phosphotransferase (*APH(3″)-Ib*), integron-encoded dihydrofolate reductase of *E. coli* (*dfrA17*), and *TEM-1*. In the UTI-2_245U strain, the following unique genetic elements were found: *sat1*, *mdtM*, *tetD*, *CMY-60*, *tetA*, *ANT (3″)-1a*, *dfrA1*, and *tetR*. No unique genetic elements were identified in the UTI-3_455U strain (Table 3).

The three UPEC strains were characterized by a resistome with genes for outer membrane proteins (*ompF*) associated with resistance to penicillin (*arnA*, *bacA*, *mgrB*, *phoP*, and *pmrBCF*) and lipopeptides. The genes for aminoglycoside phosphotransferase enzymes [*APH (3″)-1b*, *APH(3″)-Ib*, *ANT (3″)-1a*], and *kdpE* were associated with resistance to aminoglycosides. Moreover, the *gyrA*, *mfd*, and *parCE* genes conferred resistance to quinolones and fluoroquinolones in the UPEC strains. Resistance to fosfomycin was related to the presence of the *glpT*, *murA*, and *uhpT* genes. Likewise, the *nfsA* gene was found to be associated with resistance to nitrofurans (Table 3).

The genome analysis of the UTI-1_774U strain showed that the *TEM-1* gene encodes an ESBL, and the genome analysis of the UPEC strain UTI-2_245U showed that the *CMY-60* gene encodes a β-lactamase resistant to cephamycin. Macrolide resistance was associated with the *mphA* and *mrx* genes in the UTI-1_774U and UTI-3_455U strains. However, in the genome of the UTI-2_245U strain, genes related to its resistance to macrolides were not identified. Resistance to tetracycline antibiotics was mainly associated with the *acrR* gene in the UTI-1_774U and UTI-3_455U strains and with the *tetADF* operon in the UTI-2_455U strain. The *folP* and *leuO* genes, which confer resistance to folate pathway antagonists, were identified in the genomes of the three UPEC strains. However, only strains UTI-1_774U and UTI-3_455U presented integron-encoded dihydrofolate reductase (*dfrA17* and *dfrA1*) and *sul2* genes for resistance to folate pathway antagonists and sulfonamides. The three strains, UTI-1_774U, UTI-3_455U, and UTI-2_455U, were phenotypically and genotypically sensitive to cephems and cephalosporins, monobactams, carbapenems, glycylcyclines, and phenicols (Table 3). The Spearman correlation coefficient showed a non-relationship between the virulence genes and the resistance genes of the 172 *E. coli* sequences (Supplementary Figure S7).

3.2.6 Genetic elements associated with UPEC resistance

A comparison of the results using the CARD database showed 72 genetic elements associated with antibiotic resistance (GEAR) that were shared by the three UPEC strains included in this study (Table 3). The *sul2* gene was shared between strains UTI-1_774U and UTI-2_245U, while 5 GEARs (*mrx*, *pmrE*, *fyuA*, *mphA*, and *vgaC*) were identified in the two strains UTI-1_774U and UTI-3_455U. In addition, eight unique GEARs (*CMY-60*, *aadA*, *dfrA1*, *mdtM*, *sat1*, *tetA*, *tetD*, and *tetR*) were identified in strain UTI-2_245U, and four unique GEARs (*APH(3″)-Ib*, *APH(6)-Id*, *TEM-1*, and *dfrA17*) were identified in strain UTI-1_774U (Table 3; Figure 6A).

Additionally, a comparative analysis of the strains that displayed a GEAR (*sul2*) within their genomes revealed that they shared two plasmids, pEcoUTI1d from UTI-1_774U and the plasmid pEcoUTI2f from UTI-2_245U. A GEAR (*mphA*) was also identified between the plasmids pEcoUTI1d and pEcoUTI3e from the strains UTI-1_774U and UTI-3_455U, respectively. Four unique GEARs (*dfrA17*, *TEM-1*, *aph(3″)-Ib*, and *aph(6)-Id* in pEcoUTI1d) for UTI-1774 U and 2 GEARs (*dfrA1*, *sat2*, and *aadA1* in pEcoUTI2f) in the UTI-2_245U strain were identified (Figure 6A).

3.2.7 Antibiotic resistance plasmids in UPEC strains

With SnapGene software, four resistance-carrying plasmids were identified (pEcoUTI1d, pEco UTI3e, pEcoUTI2f, and pEcoUTI2c; Figures 6B–D). The analysis of the alignment of the region between 41,614 bp and 61,580 bp in the plasmid pEcoUTI1d of the strain UTI-1_774U and the region from 40,355 bp to 51,850 bp in the plasmid pEcoUTI3e of UTI-3_455U showed 86.2% similarity between them. These two regions were identified within a mobile element of ~40,000 bp flanked by the *lolE* and *mer* genes. The differences between these regions showed two insertions: (1) An insertion of a 3,347 bp fragment located between an IS1 insertion sequence and a TnAs3 transposase. The resistance gene *dfrA17*, a class 1 integrase, and a Tn3 resolvase were also identified in this insertion. (2) A 9,489 bp insertion fragment harboring the resistance genes *TEM-1*, *sul2*, *aph(3″)-Ib*, and *aph(6)-Ib* and the rep replication site is located between two transposases (Figure 6B).

A 17,888 bp region between the *lolE* and *mer* genes (36,865 bp to 54,573 bp) is in the plasmid pEcoUTI2f of the UTI-2_245U strain. This region was characterized by harboring the resistance genes *dfrA1, sat2, aadA1, sul2*, and a class I integrase. Several genes that encode hypothetical proteins, transposons, transcriptional regulators, and a partial fragment of the rep replication region were identified (Figure 6C). Finally, the plasmid pEcoUTI2c in the strain UTI-2_245U

TABLE 3 Phenotypic and genotypic analysis of the antibiotic resistance profile of recurrent UPEC clinical strains.

Strain	UTI-1_774U	UTI-2_245U			UTI-3_455U	
No. of elements resistance associated	84	86			80	
Efflux pumps genes	50	53			51	
Unique elements	*aph(6)-Id, aph(3")-Ib, dfrA17, tem-1*	*sat-1, mdtM, tetD, CMY-60, tetA, ant(3")-Ia, dfrA1, tetR*			None	
ESBL	Non–ESBL-producing	Non–ESBL-producing			Non–ESBL-producing	
Antimicrobial categories	**Sc**	**Genes**	**Sc**	**Genes**	**Sc**	**Genes**
Penicillin's	R	*ompF*	S	*ompF*	S	*ompF*
β-lactam combination agents	R	*dfrA17**	R	*dfrA1**	S	
Cephems		*tem-1**		*CMY-60,*		
cephalosporins II	S		S		S	
cephalosporins III	S		S		S	
cephalosporins IV	S		S		S	
Monobactams	S		S		S	
Carbapenems	S		S		S	
Lipopeptides	I	*arnA, bacA, mgrB, phoP, pmrBCEF*	I	*arnA, bacA, mgrB, phoP, pmrBCF*	I	*arnA, acA, mgrB, hoP, pmrBCEF*
Aminoglycosides	S	*APH(3")-Ib*, APH(6)-I*, fyuA, kdpE*	S	*kdpE, ANT(3")-Ia, ANT(3")-Ia**	S	*KdpE, fyuA*
Macrolides	R	*mphA*, mrx*	R		R	*mphA*, mrx*
Tetracyclines	S	*acrR**	R	*tetADR*	S	*acrR**
Glycylcyclines	S		S		S	
Quinones and fluoroquinolones	R	*gyrA, mfd, parCE*	S	*gyrA, mfd, parCE*	R	*parCE, gyrA, mfd*
Folate pathway antagonist	R	*folP, leuO, sul2**	R	*folP, leuO, sul2**	S	*folP, leuO*
Phenicols	S		S		S	
Fosfomycin	Ne	*glpT, murA, uhpT*	Ne	*glpT, murA, uhpT*	Ne	*glpT, murA, uhpT*
Nitrofurans	S	*nfsA*	S	*nfsA*	S	*nfsA0*
Profile	MDR-6	MDR-5			MDR-3	

No: number; Sc: susceptibility; R: resistant; S: sensitive; I: intermediate; Ne: non-evaluated; MDR: multidrug resistant; *: plasmidic gene. Penicillins (ampicillin); β-lactam combination agents (amoxicillin/clavulanate and piperacillin-tazobactam); Cephems [Cephalosporins 2 (cefaclor and cefoxitin), cephalosporins 3 (ceftazidime, cefotaxime, and ceftriaxone), cephalosporins 4 (cefepime)]; Monobactams (aztreonam); Carbapenems (imipenem, doripenem, ertapenem, and meropenem); Lipopeptides (colistin); Aminoglycosides (gentamicin and amikacin); Macrolides (erythromycin); Tetracyclines (tetracycline); Glycylcyclines (tigecycline); Quinolones and flouroquinolones (ciprofloxacin and ofloxacin); Folate pathway antagonist (trimethoprim-sulfamethoxazole); Phenicols (chloramphenicol); Nitrofurans (nitrofurantoin). *ANT* (aminoglycoside nucleotidyltransferase); *APH* (aminoglycoside phosphotransferase); *arnA* (UDP-4-amino-4-deoxy-L-arabinose formyltransferase); *bac* (undecaprenyl-diphosphatase); *CMY* (class C beta-lactamase cephamycins resistance-related); *dfr* (dihydrofolate reductase); *fol* (dihydropteroate synthase); *fyu* (ferric yersiniabactin uptake); *glp* (glycerol-3-phosphate); *gyr* (gyrase); *kdp* (potassium uptake regulator); *leu* (leucine response transcription factor); *mfd* (mutation frequency decline transcription-repair coupling factor); *mgr* (magnessium responsive promoter); *mph* (macrolide phosphotransferase); *mrx* (macrolide inactivation gene with unknow function); *mur* (UDP-N-acetylglucosamine 1-carboxyvinyltransferase); *nfs* (oxygen-insensitive NADPH nitroreductase); *omp* (outer membrane porin); *par* (DNA topoisomerase); *pho/pmr* (phosphoethanolamine transferase); *sul2* (sulfonamide resistant dihydropteroate synthase); *TEM,* ("Temoniera" Class A beta-lactamase); *uhp,* (hexose phosphate transport).

showed a region of 3,501 bp (393 bp to 3,894 bp), with a nonfunctional gene for the β-lactamase *CMY-136*, which is inactivated by insertion sequences. In this region, the *sugE* and *ecsC* pseudogenes were also inactivated by insertions (Figure 6D).

3.2.8 Genotypic and phenotypic analysis of the flagellum in clinical strains of UPEC

After incubation in mobility medium at 37°C in a 5% CO_2 incubator for 18 h, the UTI-2_245U strain showed a mobility halo of 65 mm in diameter, while the positive control strain UPEC CFT073 exhibited a halo of 32 mm in diameter. However, the strains UTI-1_774U and UTI-3_455U in addition to the negative control *E. coli* W3110 strain did not show motility halos (Figure 2C). The development of flagellar mobility as a function of time showed that the UTI-2_245U and CFT073 strains began to develop mobility halos after 3 h of incubation. Moreover, the qualitative mobility analysis showed a more pronounced mobility curve and a higher speed than the UPEC control strain CFT073 (Figure 2D). The TEM micrographs

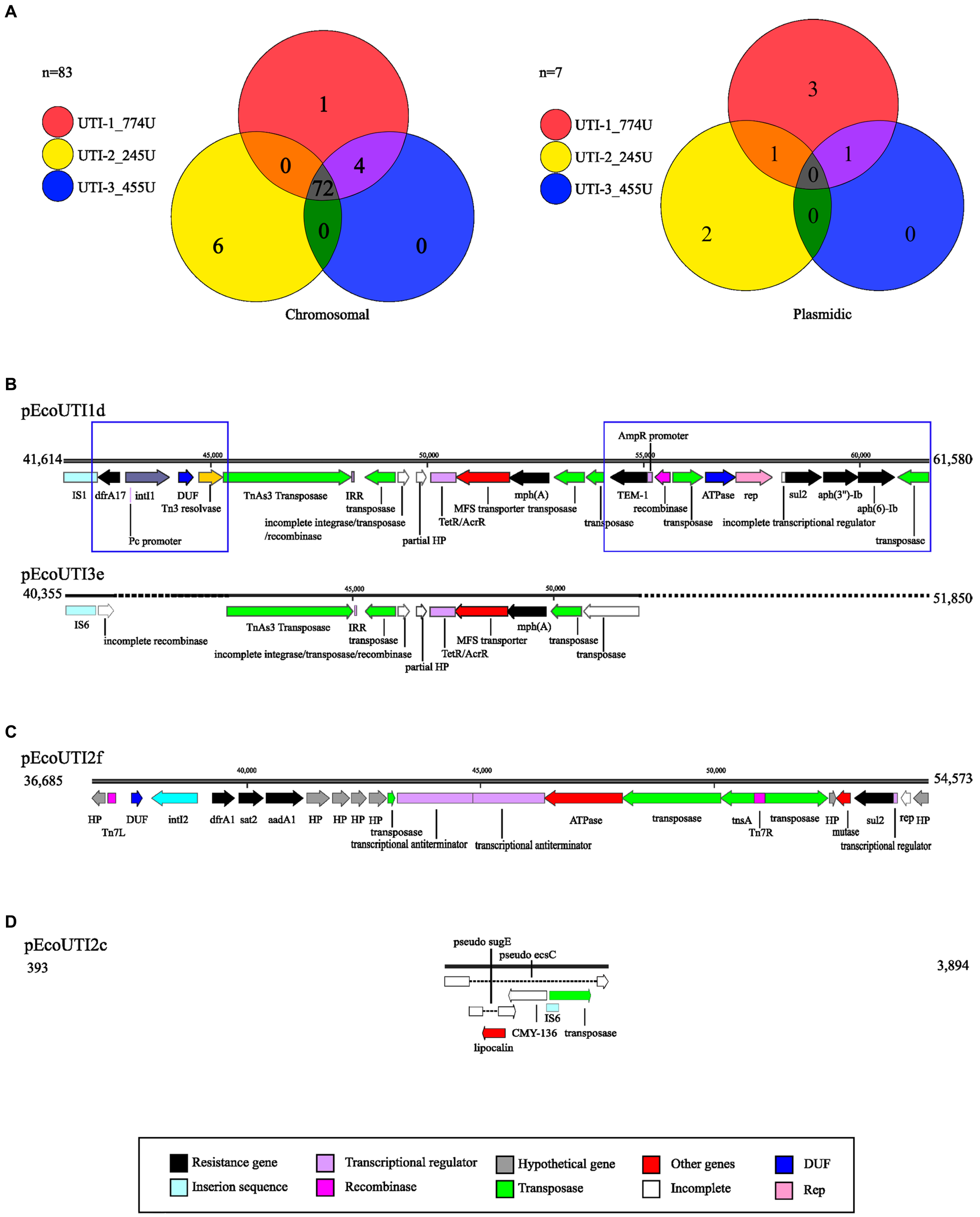


FIGURE 6
Genetic elements associated with resistance (GEAR). **(A)** Venn diagrams representing common and independent GEAR found with CARD on the chromosome and plasmids of UPEC strains. Each strain is represented with a circle of the corresponding color. **(B)** Alignment comparing the resistance-carrying region of plasmids pEcoUTI1d and pEcoUTI3e. **(C)** Representation of the resistance-carrying region of the pEcoUTI2f plasmid. **(D)** Representation of the resistance carrier region of the pEcoUTI2c plasmid. The coding sequences are shown in colors following the legend key.

of the UTI-2_245U and CFT073 strains cultured in pleuropneumonia-like organism (PPLO) medium showed long, thick, flexible, and polar structures suggestive of the presence of flagella. However, TEM micrographs of the UTI-1_774U, UTI-3_455U, and *E. coli* W3110 strains grown under the same conditions did not show these characteristic flagella structures (Figure 2E). Although bioinformatic analyses were mainly used in this study, the genomes showed the presence of genes associated with flagellar biogenesis, which are shown in Supplementary Table S1; Supplementary Figure S8.

4 Discussion

UTIs are prevalent worldwide and frequently cause outpatient and emergency medical interventions. Treating and controlling UTIs is a costly challenge for public health systems around the world, and they are one of the most common reasons for the use of broad-spectrum antibiotics (Sihra et al., 2018). The average recurrence rate of UTIs in women has been reported to be more than 30–40% (Kwok et al., 2022). The duration of an RUTI can be variable; however, in most women, recurrences occur between 6 and 12 months (Jung and Brubaker, 2019). The three patients in this study experienced several recurrence periods per year over a period of 1–3 years before undergoing a bladder biopsy. Adult women with RUTIs may present with very variable clinical symptoms, which reduce their quality of life and require individualized attention. RUTIs can occur due to an unresolved infection, reinfection, or relapse, and it is estimated that approximately 80% of RUTI cases in women occur as a result of reinfection (Pigrau-Serrallach, 2005). The clinical history recovered from the patients indicated chronic urinary infection (Chamoun et al., 2020), follicular cystitis (Mateos Blanco et al., 2007), and interstitial cystitis (Warren et al., 2008), clinical entities that may be related to RUTI episodes. The risks factors for an RUTI are essential since they said the selection of a supportive nonantibiotic treatment. Risk factors frequently implicated in the development of an RUTI have included diabetes, recent sexual intercourse, history of urogenital surgery, urinary emptying dysfunction, accidental intestinal leakage, and urinary incontinence (Sihra et al., 2018; Jung and Brubaker, 2019; Ahmed et al., 2023), as observed in the clinical data of the patients in this study. In older adult women, these comorbidities can be complicated by pyelonephritis (1.6%) or urosepsis (0.08%), conditions that put the patient's life at serious risk (Bradley et al., 2022).

In Mexico, the high incidence of obesity (42.2%) and diabetes mellitus (29%) is a political, social, and public health phenomenon that directly affects the prevalence and prognosis of RUTIs and requires new treatment strategies. Jointly, the patients in this study were subjected to topical nonantibiotic and hormone replacement treatments as an alternative to reduce symptoms and possibly eradicate the infection. These three patients received treatments such as intravesical installation of hyaluronate, oral lactobacilli, anticholinergics, topical oestrogens, and behavioral therapy. However, the alternative treatments were only effective during the application time. Various nonantibiotic management options for RUTIs have been frequently reported, with variable results depending on the progression of the RUTI, the patient's age, and the use of previous treatment (Sihra et al., 2018).

Microbiological analysis of urine cultures has been considered the "gold standard" in the diagnosis of an RUTI; however, its diagnostic accuracy has been very poor (Kwok et al., 2022). Microbiological procedures have allowed the recovery of clinical UPEC strains obtained from biopsy tissue during an RUTI episode. Bladder biopsy tissue obtained by transurethral cystoscopy has rarely been used to diagnose UTIs (Sycamore et al., 2014). However, although this invasive process has not been wholly accepted to diagnose UTIs, it has been used to diagnose interstitial cystitis (Trifillis et al., 1995; Natale et al., 2022). In this study, biopsies of bladder epithelium and urine samples were obtained from 3 patients diagnosed with RUTI for culture and isolation of UPEC strains. The presumptive evaluation of the similarity between urine strains and biopsies from the same patients was carried out by PFGE for epidemiological surveillance and the complete genome sequencing for determining the clonal relationship between strains (Blankenship et al., 2023). In this work, PFGE allowed us to observe clonality between urine and biopsy strains from the same patient. Moreover, our work is the first WGS study on UPEC strains obtained by bladder biopsy tissue from adult women with a history of RUTI for several years. The UPEC strains were analyzed with WGS molecular technology to improve diagnosis and inform the decision-making process for the treatment of patients with RUTIs. WGS has allowed the comparison of UPEC bacterial genomes from patients with complicated UTIs, uncomplicated UTIs, and RUTIs. However, even with complete genome data, it has been difficult to identify genetic characteristics associated with a specific clinical condition. This concern has been previously reported, and international organizations have made efforts to harmonize the genomics databases (Anagnostopoulos et al., 2022). WGS has facilitated the identification of common clones, clonal complexes, and essential serotypes between animals and humans, suggesting a bidirectional clonal transfer as an essential transmission method of bacteria (Flament-Simon et al., 2020). In closed environments, such as retirement homes for elderly individuals, WGS has identified RUTIs caused by the same strain of UPEC (Hidad et al., 2022). In our study, the three genomes of the UPEC strains were very closely related, observing an identity of 99% between UTI-1_774U and UTI-3_455U and an identity of 88–93% when compared with the strain UTI-2_245U. The genomic analysis of the three strains associated with RUTIs showed areas of low similarity, with loss of identity and coverage in the genes encoding pathogenicity islands (PAIs), compared to the genes encoding PAIs in the UPEC control strain CFT073. The characterization analysis of the *E. coli* strains suggests that the capacity to colonize new niches during intestinal and extraintestinal pathologies is closely associated with the presence/absence of widely recognized colonization, virulence, and fitness factors (Kaper et al., 2004). These factors are frequently grouped into PAIs and are responsible for causing a specific clinical condition in a susceptible host (Desvaux et al., 2020). Our findings suggest that RUTI strains did not present a specific genomic characteristic differentiating them from other pathotypes. We hypothesize that the loss of VFs results in the attenuation of the strains and favors the development of a chronic infection in the lower urinary tract. Furthermore, the acquisition of genes by horizontal transfer was a common element among the RUTI strains in our study, showing numerous plasmids, insertion sequences, prophages, and PAIs. The UTI-2_245U strain showed multiple CRISPR arrays, which were not related to virulence or resistance factors; however, they represented elements with inherent potential to be platforms that contribute to genomic plasticity and the horizontal acquisition of genetic material, which may have clinical relevance among different isolates (Rozwadowski and Gawel, 2022).

The comparative genomics together with the creation and curation of a database of 158 complete sequenced genomes of human urinary *E. coli* (uEc) strains allowed us to analyze our RUTI strains in evolutionary, geographical, and pathogenic contexts. The genomic analysis showed that the uEc strains are closely related (99.95% similarity), grouping into clades of intestinal and extraintestinal pathotypes; likewise, they were grouped into subclades based on their STs and serotypes (Valiatti et al., 2020; Schüroff et al., 2021). However, no relationship with the clinical conditions to which they were related was observed (Matamoros et al., 2017; Flament-Simon et al., 2020). The strains of uEc, cUTI, and RUTI in other studies were mainly associated with the O25:H4/ST131 clones, producers of ESBL and MDR that are widely distributed and cause extraintestinal infections (Contreras-Alvarado et al., 2021). The strains UTI-1_774U and UTI-3_455U were independently clustered, with the closest control genome being that of neonatal meningitis-associated *E. coli* S88 (Peigne et al., 2009), while the strain UTI-2_245U clustered together with the control genome of enterohaemorrhagic *E. coli* Sakai (Makino et al., 1999). The strains UTI-1_774U and UTI-3_455U belonged to phylogenetic group B2, serotype O75:H5/ST1193, and were resistant to fluoroquinolones. This clone has been reported in Spain, Europe, North America, and various parts of the world (García-Meniño et al., 2022; Pitout et al., 2022) as a pathogenic and MDR emerging clone in humans (probability ≥92%). This clone has been associated with community UTIs and bloodstream infections (García-Meniño et al., 2022; Pitout et al., 2022). The UPEC strains UTI-1_774U and UTI-3_455U of Mexican origin presented a set of chromosomal mutations (*gyrA* S83L, *gyrA* D87N, *parC* S80I, and *parE* L416F) that conferred resistance to fluoroquinolones. These strains presented many genomic characteristics related to RUTIs and not community UTIs similar to those reported by García-Meniño et al. (2022). The results suggest that clonal lineages related to uncomplicated and community-based UTIs may have pathogenic characteristics to adapt to recurrent urinary strains if the appropriate environmental and host factors are present. Interestingly, UTI-1_774U and UTI-3_455U were related to the clinical cUTI strain named EcPF40 in the database, reported by Sharon et al. (2020) to be associated with RUTI. Our work supports the proposal to monitor ST1193 as a new clone at the same level of importance as ST131 (Pitout et al., 2022). The UTI-2_245U strain was grouped into the phylogroup B1 and serotype O141:H5/ST295. This clone of mainly Asian origin was grouped with the controls of intestinal origin and allowed us to infer that this clade is transitional between strains from intestinal infections and extraintestinal strains (Fuentes-Castillo et al., 2023). Previously, it has been reported that slight changes in the core genome can result in changes in a strain from an intestinal environment to a highly specialized extraintestinal pathotype (McNally et al., 2013). ST295 is associated with animal hosts, giving this ST a strong zoonotic component. Fuentes-Castillo et al. (2023) reported the association of ST295 clones with seagulls; the authors conclude that the migratory routes of these animals can contribute to the dissemination of these clones in new environments (Fuentes-Castillo et al., 2023). The clones of ST295 have been identified in hospital wastewater samples, suggesting that UTI-associated bacteria play a role in emerging antibiotic-resistance genotypes in hospital settings. A high diversity of STs and phylogenetic groups with pathogenic lineages can increase hospital and community dissemination (Davidova-Gerzova et al., 2023). The discovery of clone ST295 in a patient with RUTI supports the urgent need to monitor the distribution of this clone and its correlation with other possible reservoirs.

Manual analysis of virulence and fitness factors allowed the identification of various partial, incomplete, inactive, or absent operons and regions of interest associated with adherence. The role of fimbriae as the main factors in adherence, colonization, and invasion of *E. coli* to the urothelium was of great importance in this work (Desvaux et al., 2020; Hozzari et al., 2020). The comparative genomic analysis of the fimbrial operons in 10 RUTIs genomes, 13 cUTI genomes, and 10 control genomes of intestinal and extraintestinal *E. coli* strains showed the grouping of the strains UTI-1_774U and UTI-3_455U in the clade shared by extraintestinal strains and the UTI-2_245U strain with the intestinal strains. Interestingly, the analysis revealed the absence of the *pap* operon, located in PAI I_{CFT073}, in 88.5% of the strains. P fimbriae are vital for adherence to the renal epithelium and is associated with invasive *E. coli* strains (Biggel et al., 2020). However, the analysis showed that the RUTIs strains only conserved the regulatory gene *papX*, which has participated as an inhibitor of flagellar mobility (Simms and Mobley, 2008) and whose inhibitory effect may be related to the negative result in the mobility assays with the UTI-1_774U and UTI-3_455U strains. The genomic analysis of the UTI-2_245U strain showed a loss of genes associated with type 1 fimbriae (FT1). FT1 is the most relevant classic factor in the invasion of UPEC to the urinary tract (Eto et al., 2007). Nevertheless, ablation of FT1 did not wholly abate invasion in the *in vitro* assays, and it has been observed that the ECP (*E. coli* Common Pilus) fimbriae can promote invasion in the mouse bladder (Saldaña et al., 2014). In this context, the presence of other fimbriae operons similar to FT1, such as the F9 fimbriae, which is upregulated 6-fold in cUTI (Conover et al., 2016) and F17-like, which promotes the formation of reservoirs in the intestine (Spaulding et al., 2017), may contribute to reinfections and favor the establishment of UPEC strains. Data with the UPEC reference strain CFT073 revealed the presence of 10 types of fimbriae belonging to the chaperone-accommodator family and two related to type IV pili (Welch et al., 2002). Previous studies have confirmed the role of fimbria: fimbria type 1 (*fim*) (Martinez, 2000), fimbria P (*pap*) (Kuehn et al., 1992), fimbria F1C/S (*foc/sfa*) (Khan et al., 2000), and AFA/DRA fimbria (*afa/dra*) (Selvarangan et al., 2004) in the pathogenicity of UPEC to the urinary tract. The findings in this work indicate that fimbriae 1, P, F1C/S, and AFA/DRA are rare in cUTI and RUTI strains. Miyazaki et al. (2002) reported that fimbriae 1, P, and S are not essential for UPEC colonization in the urinary tract. The results obtained in this work suggest that the high diversity of fimbrial operons and even the participation of other surface proteins, such as FimA, BtuB, ChuA, FepA, FyuA, UidC, NmpC, OmpA, OmpC, OmpF, OmpT, Flu, CarB, and MdhK are expressed during the growth of *E. coli* in urine, and in the colonization of the urinary tract (Wurpel et al., 2016). Other fimbrial operons identified with high frequency in the genomes of cUTI, RUTIs, UTI-1_774U, UTI-2_245U, and UTI-3_455U were the Curli operon (*csg*) (Luna-Pineda et al., 2019) and Ecp operon (Stacy et al., 2014), Yfc operon (Spurbeck et al., 2012), Yeh operon (Ravan and Amandadi, 2015), and F17-like operon (Martin et al., 1997), which are considered nonclassical virulence factors of urinary strains. Toxins are virulence factors associated with colonization of the urinary tract. In addition, they allow the dispersion of bacteria to tissues, promote the lysis of host cells, facilitate the release of micronutrients such as iron, which will subsequently be transported by siderophores (Dhakal and Mulvey, 2012), and increase neutrophil survival (Los et al., 2013). The *clbA* and *clbQ* genes that encode a colibactin, the *pic* gene that encodes a serine protease, the *vat* gene that encodes a hemoglobin protease, and the *hlyA* gene for α-haemolysin are factors widely described in the UPEC reference strain CFT073 (Luo et al., 2009).

The RUTI strains UTI-1_774U and UTI-3_455U did not present the exotoxins Hly, Cnf, Ast, Cdt, and Clb or the toxins Sat, Vat, and Tsh. Moreover, the UTI-2_245U strain did not present any genes associated with these toxins. As a result, we can suggest that the absence of classic UPEC VFs seems to promote the chronicity of these strains. The results allow us to hypothesize that toxins are not needed by invasive UPEC strains, mainly because they can have free access to iron in the host cell cytoplasm. Attenuation of toxicity appears to contribute to the chronic state of the infection. Similar behavior has been observed in the adaptive process of *Staphylococcus aureus* in its colonization in the nose, where various mutations promote the attenuation of cytotoxic activity and the generation of a nonhemolytic phenotype, which allows infection with reduced clinical manifestations, promoting a severe disease (Das et al., 2016). The strains UTI-1_774U, UTI-2_245U, and UTI-3_455U were shown to be resistant to penicillin, folate pathway antagonists, quinolones, and fluoroquinolones, with no resistance to cephalosporins. Furthermore, strain UTI-1_774U was MDR-6, UTI-2_245U MDR-5, and UTI-3_455U MDR-3. UPEC strains associated with RUTI have been previously reported (Luo et al., 2012; Al-Mayahie and Al Kuriashy, 2016; Issakhanian and Behzadi, 2019; Karami et al., 2021). In this study, MDR in the strains UTI-1_774U, UTI-2_245U, and UTI-3_455U was related to mobile genetic elements (MGEs). The ESBL gene, *TEM-1*, confers resistance to penicillin and cephalosporins; the genes *APH(3″)-Ib*, *APH(6)-I*, and *APH(3″)-Ia*, which confer resistance to aminoglycosides; the hydrofolate reductase genes *dfrA1* and *dfrA17*, which confer resistance to trimethoprim; the *sul2* gene, which confers sulfonamide resistance; and *the mphA* gene, which confers resistance to macrolides, were identified in plasmids of the three strains. The spread of antibiotic resistance genes (ARGs) is facilitated by mobile genetic elements (MGEs), which can be transferred from one strain to another, including integrons, transposons, and plasmids (Rozwadowski and Gawel, 2022). The RUTI strains UTI-1_774U, UTI-2_245U, and UTI-3_455U carried plasmids with mobilization genes (*mob*), transfer genes (*tra*), recombinase genes, and insertion sequence genes (IS) that facilitated their mobility; these events have been previously reported (Smillie et al., 2010). Additionally, the three strains showed a large group of small cryptic plasmids, identified by genomic homology studies between environmental strains, of veterinary importance, especially in uropathogenic clinical strains. The current function of these cryptic plasmids is unknown, and no previous reports have been found that further elucidate their importance. The evidence of the presence of repeated regions and the presence of guanine-cytosine islands suggests a possible regulatory function, and a greater number of studies are necessary to understand the role of these plasmids in RUTIs.

Furthermore, the three RUTI strains presented multiple genes for active efflux pumps and transporters that, together with the ARGs, may explain their MDR. Interestingly, the strain UTI-2_245U carried resistant genes to heavy metals. In recent years, heavy metal therapy has been proposed as an alternative to treatment in MDR strains (Wang et al., 2020). Sensitivity to bacteriophages used to treat MDR strains was also evaluated; the results indicated that the three RUTI strains were resistant to all bacteriophages evaluated (Supplementary Figure S9). Interestingly, in this work, the presence of an extrachromosomal prophage located in the UTI-2_245U strain was reported. The pEcoUTI2e plasmid belongs to the family of plasmids similar to bacteriophage the P1; within this context, this plasmid is an important source of MGEs, as has been previously reported (Billard-Pomares et al., 2014). However, in UTI-2_245U strains, this plasmid does not contain virulence or resistance factors, yet its potential for genomic plasticity is present. The phenotypic results of mobility showed the strains UTI-1_774U and UTI-3_455U as nonmotile and the strain UTI-2_245U (mobility zone 64 mm) as highly motile compared to the UPEC control strain CFT073 (mobility zone 32 mm). Genomic analysis of all flagellar genes showed that the three regions involved in flagellum biosynthesis were intact. However, minor differences were observed, including the loss of nonessential genes and the insertion of small ORFs of hypothetical proteins that do not alter the reading frame of the genes but that could be involved in the inhibition of flagellum formation (Fitzgerald et al., 2014).

The regulation of the flagellum could be influenced by the presence of a strong repressor of motility, *papX* (Simms and Mobley, 2008), present in strains UTI-1_774U and UTI-3_455U. The biofilm formation and haemolysis evaluation showed that the strains UTI-1_774U, UTI-2_245U, and UTI-3_455U were nonbiofilm producers and nonhemolytic in sheep blood. Biofilm formation is a multifactorial phenomenon that requires the presence of fimbriae, mainly type 1 fimbriae and Curli fimbriae, in addition to the flagellum, outer membrane proteins, and siderophores (Soto et al., 2007). Although Congo red uptake was evaluated as an indicator of curli generation and the strains showed fixation at 37°C (data not shown), biofilm formation could not be visualized under the conditions indicated in this study. As previously described, other relevant factors in adhesion and colonization were partially absent in the strains UTI-1_774U, UTI-2_245U, and UTI-3_455U; we suggest that the absence of one or more fitness factors causes the inhibition of biofilm formation (Qasemi et al., 2021).

Furthermore, haemolytic activity in sheep erythrocytes was not observed in the RUTI strains. The haemolysin-related genes are located on a genomic island adjacent to the pheV t-RNA in the UPEC reference strain CFT073; however, in strains UTI-1_774U and UTI-3_455U, the haemolysin region was wholly lost and replaced by the genes *irp1* and *irp2* (encodes ferric regulation genes), and *fyuA* (encodes an iron-inhibition outer membrane protein) in avian pathogenic *E. coli* strains (Tu et al., 2016). In the UTI-2_245U strain, the haemolysin genes were replaced by the *sigA* gene (encodes cytopathic protease in PAI-harboring *Shigella* sp., which contributes to intestinal fluid accumulation), and loss of the *sat* gene (encodes for a secreted proteolytic autotransporter proteins) that induces cell damage during enteroaggregative *E. coli* infection (Al-Hasani et al., 2001; Vieira et al., 2020). Additionally, the UTI-2_245U strain did not present the *irp1*, *irp2*, and *fyuA* genes, but the *chuA* gene, which encodes ferrochrome receptors, and the *kpsC* and *kpsS* genes, which are associated with capsule, were identified. Additionally, the genes *rfbA*, *rfbB*, and *rfbD* were identified that are not present in the reference strain CFT073 and that code for the biosynthesis of dTDP-rhamnose, which is the precursor of lipopolysaccharides of the cell wall. Finally, the *rpoS* gene encodes an RNA polymerase factor, which was not identified in strain CFT073, and the *sigA* gene, which is an autotransporter, was identified exclusively in strain UTI-2_245U. The siderophores and regulatory factors presented by strains UTI-1_774U, UTI-2_245U, and UTI-3_455U could contribute, in an important way, to chronicity in the urinary tract (Hunstad et al., 2005; Hryckowian and Welch, 2013; Gao et al., 2023). The absence of mobility, a critical factor in the rise of UPEC, to the upper urinary tract suggests an adaptation to the lower urinary tract. However, the UTI-2_245U strain showed increased mobility associated with its commensal origin, leading us to suggest that it is still in the adaptation process. To date, there is no effective vaccine that prevents RUTIs. In this context, regulation and effective decision-making against the indiscriminate use of antibiotics is urgently needed, given that alternatives to treatment in our strains are not viable options.

5 Limitations of the study

Although the study showed exciting results, the genotypic and phenotypic characterization was carried out only with 3 UPEC strains associated with RUTI, preventing us from generalizing the results and extrapolating them to other populations of recurrent strains. Under this same concept, it was tough to identify specific characteristics in the genomes of RUTI strains that could support the complete understanding of the persistence of the UPEC strains in the urinary tract. Another aspect of this study was that the UPEC strains were not recovered before the bladder biopsy, limiting the study of the evolution of the patient's strains. In addition, a genomic comparison of the UPEC strains of the same patients was not displayed. However, the recovery of UPEC strains from bladder biopsies is an important aspect that must continue to be studied. Finally, database platforms must be updated continuously to generate more accurate knowledge that helps study recurrent bacterial populations.

6 Conclusion

We found that the decrease in classic colonization factors in our strains of RUTIs did not affect their capacity for adherence and invasion to human bladder HTB-5 cells, showing their adaptation to this ecological niche. We showed a profile of virulence, resistance, fitness factors, and high genetic diversity of RUTIs strains that could contribute to developing a recurrent infection in the urinary tract. This variability gives these strains a different arsenal of colonization and persistence strategies. Since the absence/presence of one or more virulence factors can influence the expression of other factors, we suggest that many of these systems may be redundant and cumulative. However, the expression of classical and highly virulent factors may allow UPEC strains to evade the acute immune response and promote a chronic infection in patients with a clinical predisposition to develop RUTIs. The results of this study did not allow us to identify a single factor responsible for the recurrent phenotype and genotype in UPEC strains associated with RUTIs.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

Ethical approval was not required for the studies involving humans because as part a retrospective analysis, of storage remanents of biopsy samples, that are taken as part of the clinical diagnosis as otherwise be discarted, in this protocol the tissue samples was obtain after a appropiated diagnosis of mophological abnormalities, and medica consent includes the secundary use of tissue leftovers. Non genetic data was obtain of human samples, and all sensitive personal data was eliminated of this study. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MF-O: Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Visualization. SO: Conceptualization, Formal analysis, Funding acquisition, Resources, Supervision, Writing – original draft. AC-C: Funding acquisition, Resources, Supervision, Writing – review & editing. RC-T: Investigation, Methodology, Resources, Writing – review & editing. EM-P: Formal analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. DR-B: Investigation, Methodology, Resources, Visualization, Writing – review & editing. SZ-V: Methodology, Writing – review & editing. RH-C: Funding acquisition, Resources, Writing – review & editing. MF-E: Resources, Writing – review & editing. JA-G: Resources, Writing – review & editing. DV: Conceptualization, Resources, Writing – review & editing. JX-C: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study described in this manuscript was supported by Public Federal Funds grant numbers HIM/2018/072 SSA. 1527, HIM/2021/030 SSA. 1730, HIM/2021/031 SSA. 1720, and HIM/2022/014 SSA. 1777 from the HIMFG. This work was also supported by the "Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT)" with number grant 319563 at the "Announcement for Basic Science and/or Frontier Science, Modality: Paradigms and Controversies of Science 2022".

Acknowledgments

MF-O conducted this study to fulfill the requirements of Programa de Doctorado en Ciencias Biomédicas of Universidad Nacional Autónoma de México, and received a doctoral scholarship from Consejo Nacional de Humanidades Ciencias y Tecnologías (838589). We especially thank PhD - Associate Professor Luis Y. Kameyama Kawabe (CINVESTAV-Zacatenco- IPN) from the Genetics and Molecular Biology Department for his support and María de Lourdes Rojas Morales (LANSE-CINVESTAV-IPN) for her technical assistance in the electron microscopy experiments. The clinical data were derived from the thesis of RT for graduation from the Urology and Gynecology Speciality of the Army and Air Force of Mexico.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1340427/full#supplementary-material

References

Acuña-Ruíz, A. M., and Molina-Torres, F. A. (2022). Associated epidemiologic factors with recurring infection of the lower urinary tract in pregnant women. *Rev. Med. Inst. Mex. Seguro Soc.* 4, 411–417.

Ahmed, A. E., Abdelkarim, S., Zenida, M., Baiti, M. A. H., Alhazmi, A. A. Y., Alfaifi, B. A. H., et al. (2023). Prevalence and associated risk factors of urinary tract infection among diabetic patients: a cross-sectional study. *Healthcare* 11:861. doi: 10.3390/healthcare11060861

Alcántar-Curiel, M. D., Alpuche-Aranda, C. M., Varona-Bobadilla, H. J., Gayosso-Vázquez, C., Jarillo-Quijada, M. D., Frías-Mendivil, M., et al. (2015). Risk factors for extended-spectrum b-lactamases-producing *Escherichia coli* urinary tract infections in a tertiary hospital. *Salud Pública Méx.* 57, 412–418. doi: 10.21149/spm.v57i5.7621

Al-Hasani, K., Adler, B. E. N., Rajakumar, K., and Sakellaris, H. (2001). Distribution and structural variation of the she pathogenicity island in enteric bacterial pathogens. *J. Med. Microbiol.* 50, 780–786. doi: 10.1099/0022-1317-50-9-780

Al-Mayahie, S., and Al Kuriashy, J. J. (2016). Distribution of ESBLs among *Escherichia coli* isolates from outpatients with recurrent UTIs and their antimicrobial resistance. *J. Infect. Dev. Ctries.* 10, 575–583. doi: 10.3855/jidc.6661

Anagnostopoulos, A. V., Bello, S. M., Blake, J. A., Blodgett, O., Bult, C. J., Christie, K. R., et al. (2022). Harmonizing model organism data in the Alliance of genome resources. *Genetics* 220:22. doi: 10.1093/genetics/iyac022

Anger, J., Lee, U., Ackerman, A. L., Chou, R., Chughtai, B., Clemens, J. Q., et al. (2019). Recurrent Uncomplicated urinary tract infections in women: AUA/CUA/SUFU guideline. *J. Urol.* 202, 282–289. doi: 10.1097/JU.0000000000000296

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387

Baldy-Chudzik, K., Bok, E., and Mazurek, J. (2015). Well-known and new variants of pathogenic *Escherichia coli* as a consequence of the plastic genome. *Postepy Hig. Med. Dosw.* 69, 345–361. doi: 10.5604/17322693.1145173

Ballesteros-Monrreal, M. G., Arenas-Hernández, M. M. P., Barrios-Villa, E., Juarez, J., Álvarez-Ainza, M. L., Taboada, P., et al. (2021). Morphotypes as important trait for Uropathogenic *E. coli* diagnostic; a virulence-phenotype-phylogeny study. *Microorganisms.* 9:2381. doi: 10.3390/microorganisms9112381

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., and Clermont, O. (2018). ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb. Genom.* 4:e000192. doi: 10.1099/mgen.0.000192

Behzadi, P. (2020). Classical chaperone-usher (CU) adhesive fimbriome: uropathogenic *Escherichia coli* (UPEC) and urinary tract infections (UTIs). *Folia Microbiol.* 65, 45–65. doi: 10.1007/s12223-019-00719-x

Behzadi, P., García-Perdomo, H. A., Autrán Gómez, A. M., Pinheiro, M., and Sarshar, M. (2023). Editorial: Uropathogens, urinary tract infections, the host-pathogen interactions and treatment. *Front. Microbiol.* 14:1183236. doi: 10.3389/fmicb.2023.1183236

Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., et al. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45, W30–W35. doi: 10.1093/nar/gkx343

Biggel, M., Xavier, B. B., Johnson, J. R., Nielsen, K. L., Frimodt-Møller, N., Matheeussen, V., et al. (2020). Horizontally acquired papGII-containing pathogenicity islands underlie the emergence of invasive uropathogenic *Escherichia coli* lineages. *Nat. Commun.* 11:5968. doi: 10.1038/s41467-020-19714-9

Billard-Pomares, T., Fouteau, S., Jacquet, M. E., Roche, D., Barbe, V., Castellanos, M., et al. (2014). Characterization of a P1-like bacteriophage carrying an SHV-2 extended-spectrum β-lactamase from an *Escherichia coli* strain. *Antimicrob. Agents Chemother.* 58, 6550–6557. doi: 10.1128/aac.03183-14

Blankenship, H. M., Dietrich, S. E., Burgess, E., Wholehan, J., Soehnlen, M., and Manning, S. D. (2023). Whole-genome sequencing of Shiga toxin-producing *Escherichia coli* for characterization and outbreak investigation. *Microorganisms* 11:1298. doi: 10.3390/microorganisms11051298

Bradley, M. S., Ford, C., Stagner, M., Handa, V., and Lowder, J. (2022). Incidence of urosepsis or pyelonephritis after uncomplicated urinary tract infection in older women. *Int. Urogynecol. J.* 33, 1311–1317. doi: 10.1007/s00192-022-05132-6

Brown, E. D., and Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature* 529, 336–343. doi: 10.1038/nature17042

Burnett, L. A., Hochstedler, B. R., Weldon, K., Wolfe, A. J., and Brubaker, L. (2021). Recurrent urinary tract infection: association of clinical profiles with urobiome composition in women. *Neurourol. Urodyn.* 40, 1479–1489. doi: 10.1002/nau.24707

Buxton, R. (2005). Blood agar plates and hemolysis protocols. *American Society for Microbiology*, 15.

Chamoun, M. N., Sullivan, M. J., Goh, K. G. K., Acharya, D., Ipe, D. S., Katupitiya, L., et al. (2020). Restriction of chronic *Escherichia coli* urinary tract infection depends upon T cell-derived interleukin-17, a deficiency of which predisposes to flagella-driven bacterial persistence. *FASEB J.* 34, 14572–14587. doi: 10.1096/fj.202000760r

Clermont, O., Dixit, O. V., Vangchhia, B., Condamine, B., Dion, S., Bridier-Nahmias, A., et al. (2019). Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* 21, 3107–3117. doi: 10.1111/1462-2920.14713

CLSI. (2022). *Performance standards for antimicrobial susceptibility testing; twenty-six informational, supplement. M100-S26.* Wayne, PA: Clinical and Laboratory Standards Institute.

Conover, M. S., Ruer, S., Taganna, J., Kalas, V., De Greve, H., Pinkner, J. S., et al. (2016). Inflammation-induced adhesin-receptor interaction provides a fitness advantage to uropathogenic *E. coli* during chronic infection. *Cell Host Microbe* 20, 482–492. doi: 10.1016/j.chom.2016.08.013

Contreras-Alvarado, L. M., Zavala-Vega, S., Cruz-Córdova, A., Reyes-Grajeda, J. P., Escalona-Venegas, G., Flores, V., et al. (2021). Molecular epidemiology of multidrug-resistant uropathogenic *Escherichia coli* O25b strains associated with complicated urinary tract infection in children. *Microorganisms* 9:2299. doi: 10.3390/microorganisms9112299

Das, S., Lindemann, C., Young, B. C., Muller, J., Österreich, B., Ternette, N., et al. (2016). Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. *Proc. Natl. Acad. Sci. U. S. A.* 113, E3101–E3110. doi: 10.1073/pnas.1520255113

Davidova-Gerzova, L., Lausova, J., Sukkar, I., Nesporova, K., Nechutna, L., Vlkova, K., et al. (2023). Hospital and community wastewater as a source of multidrug-resistant ESBL-producing *Escherichia coli*. *Front. Cell. Infect. Microbiol.* 13:1184081. doi: 10.3389/fcimb.2023.1184081

De Nisco, N. J., Neugent, M., Mull, J., Chen, L., Kuprasertkul, A., de Souza Santos, M., et al. (2019). Direct detection of tissue-resident bacteria and chronic inflammation in the bladder wall of postmenopausal women with recurrent urinary tract infection. *J. Mol. Biol.* 431, 4368–4379. doi: 10.1016/j.jmb.2019.04.008

Desvaux, M., Dalmasso, G., Beyrouthy, R., Barnich, N., Delmas, J., and Bonnet, R. (2020). Pathogenicity factors of genomic islands in intestinal and extraintestinal *Escherichia coli*. *Front. Microbiol.* 11:2065. doi: 10.3389/fmicb.2020.02065

Dhakal, B. K., and Mulvey, M. A. (2012). The UPEC pore-forming toxin α-hemolysin triggers proteolysis of host proteins to disrupt cell adhesion, inflammatory, and survival pathways. *Cell Host Microbe* 11, 58–69. doi: 10.1016/j.chom.2011.12.003

Eto, D. S., Jones, T. A., Sundsbak, J. L., and Mulvey, M. A. (2007). Integrin-mediated host cell invasion by type 1–piliated uropathogenic *Escherichia coli*. *PLoS Pathog.* 3:e100. doi: 10.1371/journal.ppat.0030100

Fitzgerald, D. M., Bonocora, R. P., and Wade, J. T. (2014). Comprehensive mapping of the *Escherichia coli* flagellar regulatory network. *PLoS Genet.* 10:e1004649. doi: 10.1371/journal.pgen.1004649

Flament-Simon, S. C., de Toro, M., García, V., Blanco, J. E., Blanco, M., Alonso, M. P., et al. (2020). Molecular characteristics of extraintestinal pathogenic *E. coli* (ExPEC), uropathogenic *E. coli* (UPEC), and multidrug resistant *E. coli* isolated from healthy dogs in Spain. Whole genome sequencing of canine ST372 isolates and comparison with human isolates causing extraintestinal infections. *Microorganisms* 8:1712. doi: 10.3390/microorganisms8111712

Flores-Mireles, A. L., Walker, J. N., Caparon, M., and Hultgren, S. J. (2015). Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* 13, 269–284. doi: 10.1038/nrmicro3432

Flores-Oropeza, M. A. (2017). *Adherencia y formación de biopelículas en cepas de Escherichia coli uropatogénica (UPEC) aisladas de pacientes pediátricos del Hospital Infantil de México Federico Gómez*. Bachelor Thesis (Unpublished data). Químico Bacteriólogo Parasitólogo, Escuela Nacional de Ciencias Biologicas, Instituto Politécnico Nacional, México.

Foxman, B. (2010). The epidemiology of urinary tract infection. *Nat. Rev. Urol.* 7, 653–660. doi: 10.1038/nrurol.2010.190

Fuentes-Castillo, D., Castro-Tardón, D., Esposito, F., Neves, I., Rodrigues, L., Fontana, H., et al. (2023). Genomic evidences of gulls as reservoirs of critical priority CTX-M-producing *Escherichia coli* in Corcovado gulf, Patagonia. *Sci. Total Environ.* 874:162564. doi: 10.1016/j.scitotenv.2023.162564

Gao, M., Zhao, T., Zhang, C., Li, P., Wang, J., Han, J., et al. (2023). Ferritinophagy-mediated iron competition in RUTIs: tug-of-war between UPEC and host. *Biomed. Pharmacother.* 163:114859. doi: 10.1016/j.biopha.2023.114859

García-Meniño, I., Lumbreras, P., Lestón, L., Álvarez-Álvarez, M., García, V., Hammerl, J. A., et al. (2022). Occurrence and genomic characterization of clone ST1193 Clonotype 14-64 in uncomplicated urinary tract infections caused by *Escherichia coli* in Spain. *Microbiol. Spectr.* 10:e0004122. doi: 10.1128/spectrum.00041-22

Gomez, J., Gómez-Lus, M. L., Bas, P., Ramos, C., Cafini, F., Maestre, J. R., et al. (2013). Biofilm score: is it a differential element within gram negative bacilli? *Rev. Esp. Quimioter.* 26, 97–102.

Habibi, A., and Khameneie, M. K. (2016). Antibiotic resistance properties of uropathogenic *Escherichia coli* isolated from pregnant women with history of recurrent urinary tract infections. *Trop. J. Pharm. Res.* 15, 1745–1750. doi: 10.4314/tjpr.v15i8.21

Haddad, J. M., Ubertazzi, E., Cabrera, O. S., Medina, M., Garcia, J., Rodriguez-Colorado, S., et al. (2020). Latin American consensus on uncomplicated recurrent urinary tract infection—2018. *Int. Urogynecol. J.* 31, 35–44. doi: 10.1007/s00192-019-04079-5

Hidad, S., van der Putten, B., van Houdt, R., Schneeberger, C., and Kuil, S. D. (2022). Recurrent *E. coli* urinary tract infections in nursing homes: insight in sequence types and antibiotic resistance patterns. *Antibiotics* 11:1638. doi: 10.3390/antibiotics11111638

Hozzari, A., Behzadi, P., Kerishchi Khiabani, P., Sholeh, M., and Sabokroo, N. (2020). Clinical cases, drug resistance, and virulence genes profiling in Uropathogenic *Escherichia coli*. *J. Appl. Genet.* 61, 265–273. doi: 10.1007/s13353-020-00542-y

Hryckowian, A. J., and Welch, R. A. (2013). RpoS contributes to phagocyte oxidase-mediated stress resistance during urinary tract infection by *Escherichia coli* CFT073. *MBio* 4, e00023–e00013. doi: 10.1128/mbio.00023-13

Hunstad, D. A., Justice, S. S., Hung, C. S., Lauer, S. R., and Hultgren, S. J. (2005). Suppression of bladder epithelial cytokine responses by uropathogenic *Escherichia coli*. *Infect. Immun.* 73, 3999–4006. doi: 10.1128/iai.73.7.3999-4006.2005

Issakhanian, L., and Behzadi, P. (2019). Antimicrobial agents and urinary tract infections. *Curr. Pharm. Des.* 25, 1409–1423. doi: 10.2174/138161282599919061913021 6

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* 53, 2410–2426. doi: 10.1128/JCM.00008-15

Johnson, J. R., and Russo, T. A. (2018). Molecular epidemiology of extraintestinal pathogenic *Escherichia coli*. *EcoSal Plus* 8:17. doi: 10.1128/ecosalplus.esp-0004-2017

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Jung, C., and Brubaker, L. (2019). The etiology and management of recurrent urinary tract infections in postmenopausal women. *Climacteric* 22, 242–249. doi: 10.1080/13697137.2018.1551871

Kaper, J., Nataro, J., and Mobley, H. (2004). Pathogenic Escherichia coli. *Nat Rev Microbiol* 2, 123–140. doi: 10.1038/nrmicro818

Karami, N., Sriram, K. K., Yazdanshenas, S., Lin, Y. L., Jaén-Luchoro, D., Ekedahl, E., et al. (2021). Identity of Bla CTX-M carrying plasmids in Sequential ESBL-*E. coli* isolates from patients with recurrent urinary tract infections. *Microorganisms* 9:1138. doi: 10.3390/microorganisms9061138

Kass, E. H. (1956). Asymptomatic infections of the urinary tract. *Trans. Assoc. Am. Phys.* 69, 56–64.

Khan, A. S., Kniep, B., Oelschlaeger, T. A., Van Die, I., Korhonen, T., and Hacker, J. R. (2000). Receptor structure for F1C fimbriae of uropathogenic *Escherichia coli*. *Infect. Immun.* 68, 3541–3547. doi: 10.1128/iai.68.6.3541-3547.2000

Khonsari, M. S., Behzadi, P., and Foroohi, F. (2021). The prevalence of type 3 fimbriae in Uropathogenic *Escherichia coli* isolated from clinical urine samples. *Meta Gene* 28:100881. doi: 10.1016/j.mgene.2021.100881

Kim, A., Ahn, J. H., Choi, W. S., Park, H. K., Kim, S., Paick, S. H., et al. (2021). What is the cause of recurrent urinary tract infection? Contemporary microscopic concepts of pathophysiology. *Int. Neurourol. J.* 25, 192–201. doi: 10.5213/inj.2040472.236

Kuehn, M., Heuser, J., Normark, S., and Hultgren, S. J. (1992). P pili in uropathogenic E. coli are composite fibres with distinct fibrillar adhesive tips. *Nature* 356, 252–255. doi: 10.1038/356252a0

Kwok, M., McGeorge, S., Mayer-Coverdale, J., Graves, B., Paterson, D. L., Harris, P. N. A., et al. (2022). Guideline of guidelines: management of recurrent urinary tract infections in women. *BJU Int.* 130, 11–22. doi: 10.1111/bju.15756

Lloyd, A. L., Rasko, D. A., and Mobley, H. L. (2007). Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. *J. Bacteriol.* 189, 3532–3546. doi: 10.1128/JB.01744-06

Los, F. C. O., Randis, T. M., Aroian, R. V., and Ratner, A. J. (2013). Role of pore-forming toxins in bacterial infectious diseases. *Microbiol. Mol. Biol. Rev.* 77, 173–207. doi: 10.1128/mmbr.00052-12

Luna-Pineda, V. M., Moreno-Fierros, L., Cázares-Domínguez, V., Ilhuicatzi-Alvarado, D., Ochoa, S. A., Cruz-Córdova, A., et al. (2019). Curli of uropathogenic *Escherichia coli* enhance urinary tract colonization as a fitness factor. *Front. Microbiol.* 10:2063. doi: 10.3389/fmicb.2019.02063

Luo, C., Hu, G. Q., and Zhu, H. (2009). Genome reannotation of *Escherichia coli* CFT073 with new insights into virulence. *BMC Genomics* 10:552. doi: 10.1186/1471-2164-10-552

Luo, Y., Ma, Y., Zhao, Q., Wang, L., Guo, L., Ye, L., et al. (2012). Similarity and divergence of phylogenies, antimicrobial susceptibilities, and virulence factor profiles of *Escherichia coli* isolates causing recurrent urinary tract infections that persist or result from reinfection. *J. Clin. Microbiol.* 50, 4002–4007. doi: 10.1128/JCM.02086-12

Magiorakos, A. P., Srinivasan, A., Carey, R. B., Carmeli, Y., Falagas, M. E., Giske, C. G., et al. (2012). Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* 18, 268–281. doi: 10.1111/j.1469-0691.2011.03570.x

Makino, K., Yokoyama, K., Kubota, Y., Yutsudo, C. H., Kimura, S., Kurokawa, K., et al. (1999). Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. *Genes Genet. Syst.* 74, 227–239. doi: 10.1266/ggs.74.227

Martin, C., Rousset, E., and De Greve, H. (1997). Human uropathogenic and bovine septicaemic *Escherichia coli* strains carry an identical F17-related adhesin. *Res. Microbiol.* 148, 55–64. doi: 10.1016/s0923-2508(97)81900-6

Martinez, J. J. (2000). Type 1 pilus-mediated bacterial invasion of bladder epithelial cells. *EMBO J.* 19, 2803–2812. doi: 10.1093/emboj/19.12.2803

Matamoros, S., Van Hattem, J. M., Arcilla, M. S., Willemse, N., Melles, D. C., Penders, J., et al. (2017). Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the mcr-1 gene indicates bacterial diversity but plasmid restriction. *Sci. Rep.* 7:15364. doi: 10.1038/s41598-017-15539-7

Mateos Blanco, J., Lallave Martín, F., Ramírez Zambrana, A., Laguna Alvarez, E., Toledo Serrano, M. J., and Parra Pérez, C. (2007). Follicular cystitis. Case report and bibliographic review. *Arch. Esp. Urol.* 60, 77–80. doi: 10.4321/s0004-06142007000100015

McNally, A., Cheng, L., Harris, S. R., and Corander, J. (2013). The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biol. Evol.* 5, 699–710. doi: 10.1093/gbe/evt038

Miyazaki, J., Ba-Thein, W., Kumao, T., Yasuoka, M. O., Akaza, H., and Hayshi, H. (2002). Type 1, P and S fimbriae, and afimbrial adhesin I are not essential for uropathogenic *Escherichia coli* to adhere to and invade bladder epithelial cells. *FEMS Immunol. Med. Microbiol.* 33, 23–26. doi: 10.1111/j.1574-695x.2002.tb00567.x

Mulvey, M. A., Schilling, J. D., and Hultgren, S. J. (2001). Establishment of a persistent *Escherichia coli* reservoir during the acute phase of a bladder infection. *Infect. Immun.* 69, 4572–4579. doi: 10.1128/IAI.69.7.4572-4579.2001

Natale, F., Campagna, G., Marturano, M., Caramazza, D., Panico, G., Vacca, L., et al. (2022). Is there a role for bladder biopsy in the diagnosis of non-hunner lesions interstitial cystitis? *Urol. Int.* 107, 257–262. doi: 10.1159/000525849

Ochoa, S. A., Cruz-Córdova, A., Luna-Pineda, V. M., Reyes-Grajeda, J. P., Cázares-Domínguez, V., Escalona, G., et al. (2016). Multidrug- and extensively drug-resistant uropathogenic *Escherichia coli* clinical strains: phylogenetic groups widely associated with integrons maintain high genetic diversity. *Front. Microbiol.* 7:2042. doi: 10.3389/fmicb.2016.02042

Ochoa, S. A., Cruz-Córdova, A., Rodea, G. E., Cázares-Domínguez, V., Escalona, G., Arellano-Galindo, J., et al. (2015). Phenotypic characterization of multidrug-resistant *Pseudomonas aeruginosa* strains isolated from pediatric patients associated to biofilm formation. *Microbiol. Res.* 172, 68–78. doi: 10.1016/j.micres.2014.11.005

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., et al. (2023). InterPro in 2022. *Nucleic Acids Res.* 51, D418–D427. doi: 10.1093/nar/gkac993

Peigne, C., Bidet, P., Mahjoub-Messai, F., Plainvert, C. L., Barbe, V. R., Médigue, C., et al. (2009). The plasmid of *Escherichia coli* strain S88 (O45: K1: H7) that causes neonatal meningitis is closely related to avian pathogenic *E. coli* plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. *Infect. Immun.* 77, 2272–2284. doi: 10.1128/iai.01333-08

Pigrau, C., and Escolà-Vergé, L. (2020). Recurrent urinary tract infections: from pathogenesis to prevention. *Med. Clín.* 155, 171–177. doi: 10.1016/j.medcle.2020.04.015

Pigrau-Serrallach, C. (2005). Recurrent urinary tract infections. *Enferm. Infecc. Microbiol. Clin.* 23, 28–39. doi: 10.1157/13091446

Pitout, J. D. D., Peirano, G., Chen, L., DeVinney, R., and Matsumura, Y. (2022). *Escherichia coli* ST1193: following in the footsteps of *E. coli* ST131. *Antimicrob. Agents Chemother.* 66:e0051122. doi: 10.1128/aac.00511-22

Qasemi, A., Rahimi, F., and Katouli, M. (2021). Genetic diversity and virulence characteristics of biofilm-producing uropathogenic *Escherichia coli. Int. Microbiol.* 25, 297–307. doi: 10.1007/s10123-021-00221-w

Ravan, H., and Amandadi, M. (2015). Analysis of yeh fimbrial gene cluster in *Escherichia coli* O157: H7 in order to find a genetic marker for this serotype. *Curr. Microbiol.* 71, 274–282. doi: 10.1007/s00284-015-0842-6

Reichhardt, C., Jacobson, A. N., Maher, M. C., Uang, J., McCrate, O. A., Eckart, M., et al. (2015). Congo red interactions with curli-producing E. Coli and native curli amyloid fibers. *PLoS One* 10:e0140388. doi: 10.1371/journal.pone.0140388

Rozwadowski, M., and Gawel, D. (2022). Molecular factors and mechanisms driving multidrug resistance in uropathogenic *Escherichia coli*—an update. *Genes* 13:1397. doi: 10.3390/genes13081397

Saeki, E. K., Yamada, A. Y., De Araujo, L. A., Anversa, L., Garcia, D. D. O., De Souza, R. L. B., et al. (2021). Subinhibitory concentrations of biogenic silver nanoparticles affect motility and biofilm formation in *Pseudomonas aeruginosa*. *Front. Cell. Infect. Microbiol.* 11:656984. doi: 10.3389/fcimb.2021.656984

Saldaña, Z., De la Cruz, M. A., Carrillo-Casas, E. M., Durán, L., Zhang, Y., Hernández-Castro, R., et al. (2014). Production of the *Escherichia coli* common pilus by uropathogenic *E. coli* is associated with adherence to Hela and HTB-4 cells and invasion of mouse bladder urothelium. *PLoS One* 9:e101200. doi: 10.1371/journal.pone.0101200

Schneider, G. R., Dobrindt, U., Brüggemann, H., Nagy, G. B., Janke, B., Blum-Oehler, G., et al. (2004). The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536. *Infect. Immun.* 72, 5993–6001. doi: 10.1128/iai.72.10.5993-6001.2004

Schüroff, P. A., Salvador, F. A., Abe, C. M., Wami, H. T., Carvalho, E., Hernandes, R. T., et al. (2021). The aggregate-forming pili (AFP) mediates the aggregative adherence of a hybrid-pathogenic *Escherichia coli* (UPEC/EAEC) isolated from a urinary tract infection. *Virulence* 12, 3073–3093. doi: 10.1080/21505594.2021.2007645

Secretaría de Salud. (2023). *Boletines Epidemiológicos Semanales*. Dirección General. Available at: https://www.gob.mx/salud/acciones-y-programas/historico-boletin-epidemiologico.

Selvarangan, R., Goluszko, P., Singhal, J., Carnoy, C., Moseley, S., Hudson, B., et al. (2004). Interaction of Dr adhesin with collagen type IV is a critical step in *Escherichia coli* renal persistence. *Infect. Immun.* 72, 4827–4835. doi: 10.1128/iai.72.8.4827-4835.2004

Sharon, B. M., Nguyen, A., Arute, A. P., Hulyalkar, N. V., Nguyen, V. H., Zimmern, P. E., et al. (2020). Complete genome sequences of seven uropathogenic *Escherichia coli* strains isolated from postmenopausal women with recurrent urinary tract infection. *Microbiol. Resour. Announc.* 9, e00700–e00720. doi: 10.1128/mra.00700-20

Sihra, N., Goodman, A., Zakri, R., Sahai, A., and Malde, S. (2018). Nonantibiotic prevention and management of recurrent urinary tract infection. *Nat. Rev. Urol.* 15, 750–776. doi: 10.1038/s41585-018-0106-x

Simms, A. N., and Mobley, H. L. T. (2008). PapX, a P fimbrial operon-encoded inhibitor of motility in uropathogenic *Escherichia coli*. *Infect. Immun.* 76, 4833–4841. doi: 10.1128/iai.00630-08

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74, 434–452. doi: 10.1128/mmbr.00020-10

Soto, S. M., Smithson, A., Martinez, J. A., Horcajada, J. P., Mensa, J., and Vila, J. (2007). Biofilm formation in uropathogenic *escherichia coli* strains: relationship with prostatitis, urovirulence factors and antimicrobial resistance. *J. Urol.* 177, 365–368. doi: 10.1016/j.juro.2006.08.081

Spaulding, C. N., Klein, R. D., Ruer, S., Kau, A. L., Schreiber, H. L., Cusumano, Z. T., et al. (2017). Selective depletion of uropathogenic *E. coli* from the gut by a FimH antagonist. *Nature* 546, 528–532. doi: 10.1038/nature22972

Spurbeck, R. R., Dinh, P. C., Walk, S. T., Stapleton, A. E., Hooton, T. M., Nolan, L. K., et al. (2012). *Escherichia coli* isolates that carry vat, fyuA, chuA, and yfcV efficiently colonize the urinary tract. *Infect. Immun.* 80, 4115–4122. doi: 10.1128/iai.00752-12

Stacy, A. K., Mitchell, N. M., Maddux, J. T., De la Cruz, M. A., Durán, L., Girón, J. A., et al. (2014). Evaluation of the prevalence and production of *escherichia coli* common pilus among avian pathogenic *E. Coli* and its role in virulence. *PLoS One* 9:e86565. doi: 10.1371/journal.pone.0086565

Stamm, W. E., and Hooton, T. M. (1993). Management of urinary tract infections in adults. *N. Engl. J. Med.* 329, 1328–1334. doi: 10.1056/NEJM199310283291808

Stepanović, S., Vuković, D., Hola, V., Bonaventura, G. D., Djukić, S., Ćirković, I., et al. (2007). Quantification of biofilm in microtiter plates: overview of testing conditions and practical recommendations for assessment of biofilm production by staphylococci. *APMIS* 115, 891–899. doi: 10.1111/j.1600-0463.2007.apm_630.x

Sycamore, K. F., Poorbaugh, V. R., Pullin, S. S., and Ward, C. R. (2014). Comparison of urine and bladder or urethral mucosal biopsy culture obtained by transurethral cystoscopy in dogs with chronic lower urinary tract disease: 41 cases (2002 to 2011). *J. Small Anim. Pract.* 55, 364–368. doi: 10.1111/jsap.12225

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H., et al. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J. Clin. Microbiol.* 33, 2233–2239. doi: 10.1128/jcm.33.9.2233-2239.1995

Thanert, R., Reske, K. A., Hink, T., Wallace, M. A., Wang, B., Schwartz, D. J., et al. (2019). Comparative genomics of antibiotic-resistant uropathogens implicates three routes for recurrence of urinary tract infections. *MBio* 10, e01977–e01919. doi: 10.1128/mBio.01977-19

The UniProt Consortium (2023). UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531. doi: 10.1093/nar/gkac1052

Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P. A., Murphy, T. D., et al. (2016). P8008 the NCBI eukaryotic genome annotation pipeline. *J. Anim. Sci.* 94:184. doi: 10.2527/jas2016.94supplement4184x

Trifillis, A. L., Cui, X., Jacobs, S., and Warren, J. W. (1995). Culture of bladder epithelium from cystoscopic biopsies of patients with interstitial cystitis. *J. Urol.* 153, 243–248. doi: 10.1097/00005392-199501000-00085

Tu, J., Xue, T., Qi, K., Shao, Y., Huang, B., Wang, X., et al. (2016). The irp2 and fyuA genes in high Pathogenicity Islands are involved in the pathogenesis of infections caused by avian pathogenic *Escherichia coli* (APEC). *Pol. J. Vet. Sci.* 19, 21–29. doi: 10.1515/pjvs-2016-0004

Ulett, G. C., and Schembri, M. A. (2016). Bacterial pathogenesis: Remodelling recurrent infection. *Nat. Microbiol.* 2, 16256–16252. doi: 10.1038/nmicrobiol.2016.256

Valiatti, T. B., Santos, F. F., Santos, A. C., Nascimento, J. A., Silva, R. M., Carvalho, E., et al. (2020). Genetic and virulence characteristics of a hybrid atypical enteropathogenic and uropathogenic *Escherichia coli* (aEPEC/UPEC) strain. *Front. Cell. Infect. Microbiol.* 10:492. doi: 10.3389/fcimb.2020.00492

Vieira, P. C., Espinoza-Culupu, A. O., Nepomuceno, R., Alves, M. R., Lebrun, I., Elias, W. P., et al. (2020). Secreted autotransporter toxin (sat) induces cell damage during enteroaggregative *Escherichia coli* infection. *PLoS One* 15:e0228959. doi: 10.1371/journal.pone.0228959

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55, 641–658. doi: 10.1373/clinchem.2008.112789

Wagenlehner, F., Wullt, B., Ballarini, S., Zingg, D., and Naber, K. G. (2018). Social and economic burden of recurrent urinary tract infections and quality of life: a patient web-based study (GESPRIT). *Expert Rev. Pharmacoecon. Outcomes Res.* 18, 107–117. doi: 10.1080/14737167.2017.1359543

Wagenlehner, F. M. E., Bjerklund Johansen, T. E., Cai, T., Koves, B., Kranz, J., Pilatz, A., et al. (2020). Epidemiology, definition and treatment of complicated urinary tract infections. *Nat Rev Urol*, 586–600. doi: 10.1038/s41585-020-0362-4

Wang, M., Goh, Y. X., Tai, C., Wang, H., Deng, Z., and Ou, H. Y. (2022). Vrprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. *Nucleic Acids Res.* 50, W768–W773. doi: 10.1093/nar/gkac321

Wang, P., Wang, J., Xie, Z., Zhou, J., Lu, Q., Zhao, Y., et al. (2020). Depletion of multidrug-resistant uropathogenic *Escherichia coli* BC1 by ebselen and silver ion. *J. Cell. Mol. Med.* 24, 13139–13150. doi: 10.1111/jcmm.15920

Warren, J. W., Brown, V., Jacobs, S., Horne, L., Langenberg, P., and Greenberg, P. (2008). Urinary tract infection and inflammation at onset of interstitial cystitis/painful bladder syndrome. *Urology* 71, 1085–1090. doi: 10.1016/j.urology.2007.12.091

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 17020–17024. doi: 10.1073/pnas.252529799

Wurpel, D. J., Totsika, M., Allsopp, L. P., Webb, R. I., Moriel, D. G., and Schembri, M. A. (2016). Comparative proteomics of uropathogenic *Escherichia coli* during growth in human urine identify UCA-like (UCL) fimbriae as an adherence factor involved in biofilm formation and binding to uroepithelial cells. *J. Proteome* 131, 177–189. doi: 10.1016/j.jprot.2015.11.001

Yang, X., Sha, K., Xu, G., Tian, H., Wang, X., Chen, S., et al. (2016). Subinhibitory concentrations of allicin decrease uropathogenic *Escherichia coli* (UPEC) biofilm formation, adhesion ability, and swimming motility. *Int. J. Mol. Sci.* 17:979. doi: 10.3390/ijms17070979

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347–W352. doi: 10.1093/nar/gkr485

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 29 January 2024
DOI 10.3389/fmicb.2024.1295184

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Wentao Zhu,
Capital Medical University, China
Fabricio Souza Campos,
Federal University of Rio Grande do Sul, Brazil

*CORRESPONDENCE
Kai Sun
✉ 41940151@qq.com
Feng Hu
✉ hf3021@shtrhospital.com
Changhui Wang
✉ wang-chang-hui@hotmail.com
Shuanshuan Xie
✉ xieshuanshuan@aliyun.com

†These authors have contributed equally to this work

RECEIVED 20 September 2023
ACCEPTED 02 January 2024
PUBLISHED 29 January 2024

CITATION
Yao A, Wang J, Xu Q, Shah BK, Sun K, Hu F, Wang C and Xie S (2024) Higher diagnostic value of metagenomic next-generation sequencing in acute infection than chronic infection: a multicenter retrospective study. *Front. Microbiol.* 15:1295184.
doi: 10.3389/fmicb.2024.1295184

Higher diagnostic value of metagenomic next-generation sequencing in acute infection than chronic infection: a multicenter retrospective study

Anjie Yao[1†], Jiale Wang[2†], Qintao Xu[1,2,3†], Binay Kumar Shah[2], Kai Sun[4*], Feng Hu[5*], Changhui Wang[1*] and Shuanshuan Xie[1*]

[1]Department of Respiratory Medicine, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [2]School of Medicine, Tongji University, Shanghai, China, [3]College of Medicine, Jinggangshan University, Ji'an, China, [4]Department of Respiratory Medicine, ChongMing Branch of Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [5]Department of Respiratory and Critical Care Medicine, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Background: The aim of this study is to compare the diagnostic value of metagenomic next-generation sequencing (mNGS) vs. conventional culture methods (CM) in chronic infection and acute infection.

Methods: We retrospectively analyzed the bronchoalveolar lavage fluid (BALF) of 88 patients with acute infection and 105 patients with chronic infection admitted to three hospitals from 2017 to 2022.

Results: The results showed that the sensitivity and specificity of mNGS were higher than those of CM. The number of patients who changed the antibiotic treatment in the mNGS positive group was larger than that of patients in the mNGS negative group in both the acute infection group (60.5 vs. 28.0%, $P = 0.0022$) and chronic infection group (46.2 vs. 22.6%, $P = 0.01112$). High levels of temperature (OR: 2.02, 95% CI: 1.18–3.70, *P*: 0.015), C-reactive protein (CRP) (OR: 15, 95% CI: 2.74–280.69, *P*: 0.011), neutrophil count (OR: 3.09, 95% CI: 1.19–8.43, *P*: 0.023), and low levels of lymphocyte count (OR: 3.43, 95% CI:1.26–10.21, *P*: 0.020) may lead to positive mNGS results in the acute infection group while no significant factor was identified to predict positive results in the chronic infection group.

Conclusion: mNGS could provide useful guidance on antibiotic strategies in infectious diseases and may be more valuable for the diagnosis and treatment of acute infection vs. chronic infection.

KEYWORDS

metagenomic next-generation sequencing, bronchoalveolar lavage fluid, acute infection, chronic infection, diagnostic

Introduction

Respiratory tract infection represents the most prevalent infectious disease and constitutes a formidable clinical challenge due to its diverse etiologies. Clinically, ∼19–62% of respiratory tract infection cases are etiologically unclear. Whether respiratory-borne infections were due to bacteria, viruses, or fungi, they have witnessed a distressing surge in both their incidence and the subsequent mortality on a global scale (Jin et al., 2022). Our ability to treat infections is jeopardized by antimicrobial use and resistance. The devastating result of infection emphasizes how crucial early diagnosis and effective antibiotic treatment are for diseases. Microbial culture, antigen/antibody assays, and polymerase chain reaction (PCR)-based nucleic acid detection are the primary components of conventional molecular testing for pathogen identification. Complete identification of fastidious organisms is often reserved for isolates in pure culture. However, it is a time-consuming and arduous technique that will take several days to complete (Li et al., 2020). In addition, the sensitivity of the conventional culture methods is heavily influenced by the duration of the infection and whether or not the patient has already received antibiotic treatment. Antigen/antibody assays have a narrow range of applications, and the results are often impacted by the threshold. Despite its excellent specificity and sensitivity, PCR still requires pathogen prediction in order to create the appropriate primers (Xiao et al., 2023).

The limitations of CM frequently result in delayed or incorrect diagnosis and even improper antibiotic administration. As a result, timely and precise identification of unknown pathogenic microorganisms is crucial for guiding clinical decision-making regarding diagnosis and therapy. Metagenomic next-generation sequencing (mNGS) is a new pathogen detection method with excellent effectiveness and has grown steadily in healthcare settings due to excellent effectiveness, a broad pathogen spectrum, and enhanced sensitivity. Theoretically, mNGS performs unbiased, meticulous high- throughput sequencing of the total DNA or RNA content of nearly all recognized pathogens, including bacteria, fungi, viruses, *Mycobacterium tuberculosis*, parasites, and atypical pathogens, and the sequence data obtained are then compared with databases (Gu et al., 2021). In addition, previous antibiotics have an effect on the diagnostic accuracy of mNGS compared with CM (Lv et al., 2023).

An updated report of Global Burden of Disease 2019 shows that chronic respiratory diseases are the third leading cause of death with mortality of 4.0 million and prevalence of 454.6 million cases globally (GBD 2019 Chronic Respiratory Diseases Collaborators, 2023). In recent years, the utility of mNGS for the detection and diagnosis of respiratory tract infections has been studied. However, the different impacts of mNGS of non-sterile body fluids such as bronchoalveolar lavage fluid (BALF) on the diagnosis and prognosis of patients with acute infection and chronic infection remain controversial. In this multicenter retrospective study, we explored the diagnostic value of mNGS in the early detection of microorganisms by comparing acute and chronic infections in BALF samples, hoping that the results could help diagnosis and treatment of patients with acute and chronic infections.

Methods

Study design and data collection

This multicenter retrospective study analyzed 193 patients with suspected pulmonary infection who were admitted from May 2017 to November 2022 at three hospitals: the General institute of Shanghai Tenth People's Hospital, the Chongming Branch of Shanghai Tenth People's Hospital, and the Tongren Hospital. The study was approved by the Ethics Committee of the Tenth People's Hospital of Tongji University. The recruitment process is shown in Figure 1. Exclusion criteria were: (1) patients younger than 18 years, pregnant women, and psychiatric patients; (2) patients with non- pulmonary infection as shown by imaging and other relevant tests; (3) patients with partial data missing or mNGS failure. Based on the course of disease, radiological images, and laboratory test results, the patients were classified into an acute infection group ($n = 88$) and a chronic infection group ($n = 105$). Patients in the acute infection group had pulmonary infection for less than a month and new infections on imaging of chest computed tomography (CT), and patients in the chronic group had pulmonary infection for more than a month with no presence of new infections on chest CT imaging. CM and mNGS results, patient clinical characteristics, blood test results, antibiotic treatments, length of hospital stay, survival outcomes during hospitalization, and the clinical outcome of each patient were collected and analyzed.

Culture method

BALF was collected from patients with suspected pneumonia by bronchoscopy within 72 h after admission to the hospital, and the collected BALF specimen was sent to the Department of Clinical Laboratory of Shanghai Tenth People's Hospital for mNGS and CM examination. BALF was qualified according to the following conditions: no airway secretion in the BALF, 40% recovery with more than 95% cell survival, 10% erythrocytes (excluding trauma/hemorrhagic factors), and 3–5% epithelial cells and undistorted intact smear cells (Wang et al., 2019). Blood agar, chocolate agar, and MacConkey agar plates were used for bacterial culture at 35°C and 5% CO2 concentration. Roche medium was used for mycobacterial culture, and Sabouraud Dextrose Agar was used for fungi at 37 and 25°C, respectively.

Metagenomic next-generation sequencing

BALF was collected by bronchoscopy, and mNGS was performed after admission or within 72 h of disease onset. After obtaining BALF samples, mNGS was performed. A standard operating procedure of the DNA-based mNGS method was developed for the diagnosis of pathogens. In brief, 1 ml of sample was centrifuged at 12,000 × g for 5 min to collect the pathogens and human cells. Next, 50 μl of precipitate underwent depletion of host nucleic acid using 1 U of Benzonase (Sigma) and 0.5% Tween 20 (Sigma) and incubated at 37°C for 5 min. Terminal buffer (400 μl) was added to stop the reaction. Then, the quantified unique DNA fragments (named UMSI) were spiked for each sample as an identity and internal control, which were PCR products of *Oryza sativa* 400–600 bp in length. In total, 600 μl of the mixture was transferred to new tubes containing 500 μl of ceramic beads for bead beating using a Minilys personal TGrinder H24 homogenizer (catalog number OSE- TH-01; Tiangen, China). Then, nucleic acid from 400 μl of pretreated samples was extracted

Abbreviations: mNGS, metagenomic next-generation sequencing; CM, culture methods; BALF, bronchoalveolar lavage fluid; PCR, polymerase chain reaction; PPV, Positive predictive value; NPV, Negative predictive value; PCT, procalcitonin; CRP, C- reactive protein; WBC, white blood cell; N, neutrophil; L, lymphocyte; CHD, cornary heart diseases.

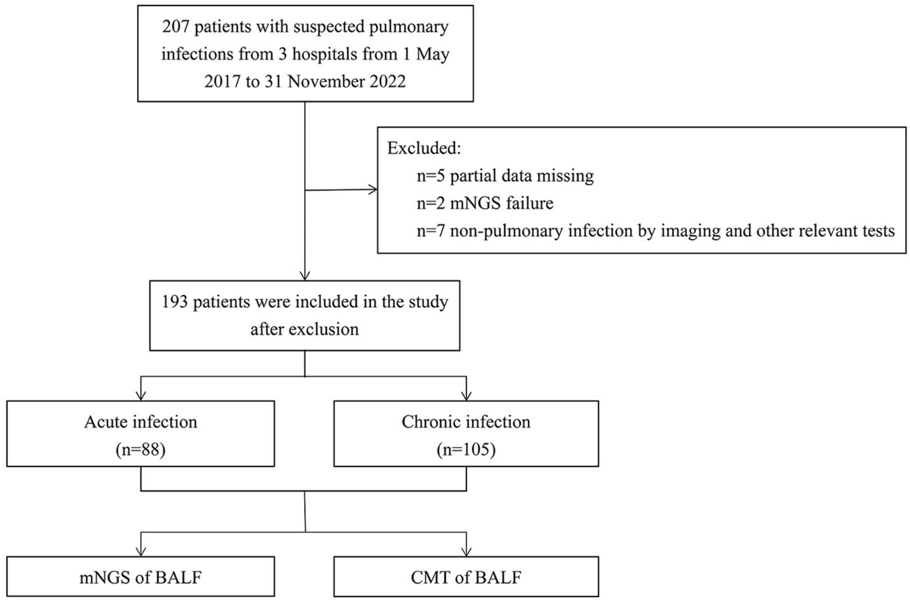

207 patients with suspected pulmonary infections from 3 hospitals from 1 May 2017 to 31 November 2022
Excluded:
n=5 partial data missing
n=2 mNGS failure
n=7 non-pulmonary infection by imaging and other relevant tests
193 patients were included in the study after exclusion
Acute infection (n=88)
Chronic infection (n=105)
mNGS of BALF
CMT of BALF

FIGURE 1
Flow diagram. mNGS, metagenomic next-generation sequencing; CMT, culture methods.

and eluted in 60 μl of elution buffer using a QIAamp UCP pathogen minikit (catalog number 50214; Qiagen, Germany). The extracted DNA was quantified using a Qubit double-stranded DNA (dsDNA) high-sensitivity (HS) assay kit (catalog number Q32854; Invitrogen, USA).

In total, 30 μl of the eluate was used to generate libraries using the Nextera DNA Flex Kit (Illumina, San Diego, CA, USA), according to the manufacturer's instructions. Fragmentation and tagmentation of the DNA were performed using the bead-linked transposome. After completion of post-tagmentation cleanup, the tagmented DNA was amplified; the thermocycling parameters were as follows: 68°C for 3 min and 98°C for 3 min, followed by 18 cycles of 45 s at 98°C, 30 s at 62°C, and 2 min at 68°C, before a final minute at 68°C. Dual indexing was conducted by employing the IDT for Illumina DNA/RNA UD indexes (catalog number 20027213). Purification and size selection were carried out following the double-sided bead purification procedure. A Qubit dsDNA HS assay kit was used to measure the library concentration. Library quality was assessed with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) using a high-sensitivity DNA kit. The library was prepared by pooling 1.5 pM concentration of each purified sample equally for sequencing on an Illumina NextSeq 550 sequencer using a 75-cycle single-end sequencing strategy.

For bioinformatics analysis, Trimmomatic was used to remove low-quality reads, adapter contamination, duplicate reads, and reads shorter than 70 bp. Low-complexity reads were removed by Kcomplexity using default parameters. The human sequence data were identified and excluded by mapping a human reference genome (hg38) using SNAP v1.0beta.18. To construct the microbial genome database, pathogens and their genomes or assemblies were selected following the Kraken2 criteria for selecting representative assemblies for microorganisms (bacteria, viruses, fungi, protozoa, and other multicellular eukaryotic pathogens) from the NCBI Assembly and Genome databases (https://benlangmead.github.io/aws-indexes/k2). Microbial reads were aligned to the database using Burrows-Wheeler Aligner software. We defined that reads with 90% identity of reference were mapped reads. In addition, reads with multiple locus alignments within the same genus were excluded from the secondary analysis. Only reads mapped to the genome within the same species were considered.

We normalized the sequencing reads RPTM to eliminate the errors caused by various sequencing depths among samples. To establish the optimal threshold value for the >10 microbes with culture isolates, samples spiked with microbes were defined as positive samples, while negative control (NC) was defined as the negative sample. Receiver operating characteristic curves were plotted for each target species using these samples. The parameter resulting in the highest area of AUC was considered the positive cutoff value for this species. For microorganisms without culture isolates, the RPTM mean value and standard deviation of this microorganism were calculated, and the RPTM [mean + 2 standard deviations (SD)] was set as a positive cutoff value.

The clinical reportable range (CRR) for pathogens was established according to the following three references indicated in a previous study: (i) the Johns Hopkins ABX Guide (https://www.hopkinsguides.com/hopkins/index/Johns_Hopkins_ABX_Guide/Pathogens), (ii) Manual of Clinical Microbiology, and (iii) clinical case reports or research articles published in peer-reviewed journals. All microbes that exceeded the threshold of mNGS were classified into three categories: (i) probable (BALF mNGS-based results were within the CRR and concordant with the clinical and radiologic results; the RPTM was significantly higher than the positive cutoff value, and the abundance was obviously higher than that of other species of the same genus), (ii) possible (the microbe has pathogenic potential, but an alternate explanation is more likely), and (iii) unlikely (the microbe cannot cause pneumonia).

To monitor the sources of potential contamination, both NC and sterile deionized water, which served as non-template controls, were prepared in parallel with other samples in each batch. In addition, we used sterile cotton swabs dipped in sterile deionized water to wipe the surfaces of the centrifuge and biosafety cabinet, to generate the background microorganism list in our laboratory.

Golden standard based on clinical compound diagnosis

Two respiratory physicians with expertise in respiratory management independently reviewed the medical records and the results of CM and mNGS of all patients. First, they determined whether the patient had a lung infection; second, they identified causative pathogens based on the patient's complaints, clinical presentation, laboratory findings, imaging presentation, and microbiological investigations (including CM and mNGS). Finally, they made a decision on the treatment regimen or adjusted the treatment regimen used. Disagreements between the two physicians regarding the causative pathogen were resolved through in-depth discussion until the consensus was reached, or if consensus could not be reached, another respiratory physician with a higher professional title would be consulted.

Criteria for mNGS positive results

The data were filtered to delete the reads with low quality, containing sequencing adapters to obtain the clean reads. After subtracting human host sequences in the clean reads, the remaining non-human sequences were compared with the microbial genome database. If the presence of a pathogenic organism met with any of the following criteria, the mNGS result was judged to be positive [RPM (reads per million) = Mapped reads number × 106/total sequencing reads; SDSMRN (stringently mapped reads number) = mapped reads number × 20 × 106/total sequencing reads]: (1) Bacteria: RPMsample/NTC $\geq$ 10, SDSMRN $\geq$3; (2) Fungi: RPMsample/NTC $\geq$ 1, SDSMRN $\geq$ 3; (3) RNA virus: RPMsample/NTC $\geq$ 1, SDSMRN $\geq$ 1; (4) DNA virus: RPMsample/NTC $\geq$ 1, SDSMRN $\geq$ 3 (Liang et al., 2023).

Statistical analysis

SPSS 25.0 and R language (4.3.1) statistical software were used for data processing and analysis and graphing. Normally distributed data were expressed as $x \pm s$; Pearson's chi-square test was used to compare the differences of categorical variables between different groups; logistic regression analysis was used to explore the risk factors associated with acute and chronic lung infections; and COX proportional risk regression was used to explore the effects of different factors on survival outcomes. All tests were two-tailed, and $P < 0.05$ was considered statistically significant.

Results

Characteristics of patients

As presented in Table 1, no significant difference was observed in the baseline characteristics between acute and chronic infection groups. The proportion of older male patients aged 60–75 years was 42.0% in the acute infection group and 46.6% in the chronic infection group. The proportion of patients complicated with other diseases (such as hypertension, diabetes, cardiac disease, stroke, and tumors) was 81.8% in the acute infection group and 65.7% in the chronic infection group, with hypertensive diseases predominating, 21.6% in the acute infection group, and 22.9% in the chronic infection group. The detection rates of CM were 22.7% in the acute infection group and 35.2% in the chronic infection group. The detection rates of mNGS were 43.2% in the acute infection group and 49.5% in the chronic infection group, which were all higher than those of CM. However, the overall detection rates were not significantly influenced by the disease characteristics in either CM or mNGS.

mNGS is superior to CM for diagnosis of infections

As shown in Table 2, mNGS was more sensitive than CM (42.3 vs. 25.5%), and its specificity was higher than that of CM (10.7 vs. 7.1%). The positive predictive values (PPV) of the two groups were 96.3 and 95.5%, and the negative predictive values (NPV) were 22.3 and 17.5%, respectively.

Additionally, there were substantial disparities in the results obtained from CM and mNGS. mNGS had the potential to significantly alter the distribution of detected pathogens, including many pathogens that could not be identified by CM. As shown in Figure 2, mNGS was dominated by bacterial infections (37%), and negative and mixed infections both accounted for 24% (Figure 2A). CM was primarily composed of negative patients, accounting for a high proportion of 67%, followed by bacterial–fungal co-infections (12%) and bacterial infections (10%) (Figure 2B). However, as presented in Figure 2C, there was a limited concordance between CM and mNGS. Only 22% of the results showed an exact match (Double- and Match), with mNGS+ alone accounting for 46%. Therefore, the distribution of bacterial and fungal detections demonstrated significant differences (Figure 3). In the acute infection group, mNGS detected a more diverse set of bacteria and fungi. CM was dominated by *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Candida albicans*, and mNGS was dominated by *Klebsiella pneumoniae*, *Prevotella*, and *Candida albicans.* There was a little difference in the distribution of bacteria in the chronic infection group, with both CM and mNGS showing dominance of *Acinetobacter baumannii, Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. However, mNGS could detect more fungi, with *Candida albicans* and *Candida glabrata* being the predominant species. In contrast, CM was dominated by *Candida albicans.*

Compared with the chronic infection group, CM exhibited a lower detection rate for bacteria in the acute infection group. However, regardless of the disease type or duration, mNGS

TABLE 1 Baseline characteristics of patients between acute infection group and chronic infection group.

Characteristics	Acute infection (n = 88)	Chronic infection (n = 105)	P-value
Sex			0.488
Female	37 (42.0%)	39 (37.1%)	
Male	51 (58.0%)	66 (62.9%)	
Age			0.362
<45	18 (20.4%)	12 (11.4%)	
45–60	22 (25.0%)	27 (25.7%)	
60–75	37 (42.0%)	49 (46.6%)	
≥75	11 (12.5%)	17 (16.1%)	
Comorbidities			0.284
No	16 (14.7%)	36 (28.1%)	
Hypertension	19 (17.4%)	24 (10.9%)	
Diabetes	12 (11.0%)	9 (7.0%)	
Coronary heart disease	8 (7.3%)	8 (6.3%)	
Arrhythmia	3 (2.8%)	2 (1.6%)	
Stroke	7 (6.4%)	6 (4.7%)	
Tumor	6 (5.5%)	7 (5.5%)	
Autoimmune disease	1 (0.9%)	4 (3.1%)	
Others	37 (33.9%)	32 (25.0%)	
CM			0.058
Negative	68 (77.3%)	68 (64.8%)	
Positive	20 (22.7%)	37 (35.2%)	
mNGS			0.379
Negative	50 (56.8%)	53 (50.5%)	
Positive	38 (43.2%)	52 (49.5%)	

mNGS, metagenomic next-generation sequencing; CM, culture methods.

TABLE 2 Comparison between mNGS and culture methods in all the patients.

Test	Sensitivity	Specificity	PPV	NPV
mNGS	47.27%	10.71%	96.3%	22.32%
CM	25.45%	7.14%	95.45%	17.45%

mNGS, metagenomic next-generation sequencing; CM, culture methods; PPV, Positive predictive value, the percentage of patients with disease determined by gold standard to the number of mNGS positive patients; NPV, negative predictive value, the percentage of patients without disease determined by gold standard to the number of mNGS negative patients.

demonstrated more consistent results and displayed a stable ability to detect bacteria, fungi, and viruses (Figure 4).

mNGS is beneficial to medication and rehabilitation

The results of mNGS are beneficial for clinicians to make decisions for antibiotics, as well as for the decline and recovery of inflammatory indicators in patients with acute infections. As shown in Table 3, in both acute and chronic infection groups, 60.5 and 46.2% of patients adjusted their antibiotic category based on the positive mNGS results, compared with those with negative results (60.5 vs. 28.0%, $P = 0.002$; 46.2 vs. 22.6%, $P = 0.011$). In addition, in the acute infection group, negative mNGS tended to predict the decline in inflammatory indicators, such as the percentage of patients with neutrophils returning to normal could be increased from 64.0 to 94.4% ($P = 0.028$); however, there was no significant benefit of mNGS in the chronic infection group for the recovery of their infection indicators. Moreover, as shown in Figure 5, the results of mNGS could not improve the survival time of the patients, and the COX regression did not show significant effects on survival about age, gender, antibiotics, mNGS, and inflammation indicators (Table 4). Therefore, the results of mNGS detection could be beneficial to clinicians to adjust the use of antibiotics against the detected pathogens and predict the recovery of inflammatory indicators in acute infections.

Recommended mNGS for patients with acute infections

We predicted the potential factors which may lead to the positive mNGS results in the acute or chronic infection groups through regression analysis. As shown in Figure 6, pathogenic microorganisms in the acute infection group were more likely to be detected by mNGS if the patients were found to have fever, C-reactive protein (CRP) > 10 mg/L, neutrophil > 75%, or lymphocyte < 20%. On the contrary, abnormalities in the above indicators in the chronic infection group could not be used to predict the positive mNGS results. In addition, as shown in Figure 7, viruses or other pathogens were more likely to be detected by mNGS in patients with coronary artery disease or arrhythmia or a high fever (T > 38°C) in the both acute and chronic infection groups (Figures 7A, C, D), whereas fungi were more likely to be detected in patients with a slight fever (37.3–38°C) in the acute infection group (Figure 7C). Furthermore, the acute infection group tended to have an increasing detection rate with fungus, especially for elderly (Figure 7B). In contrast, only viral infections were associated with elderly in the chronic infection group (Figure 7E). Therefore, mNGS was more recommended for elderly with acute infections and abnormal inflammatory markers.

Discussion

In this study, we compared the test results and clinical features of mNGS and CM for BAL samples between acute infection and chronic infection patients systematically, and the results suggest that mNGS had more advantages in some aspects as compared with CM. Firstly, mNGS is more stable, sensitive and accurate than CM in the detection of bacteria, fungi, viruses and atypical pathogens. mNGS could identify different dominant pathogens in patients with different underlying diseases and clinical characteristics in both acute and chronic infection patients. Secondly, positive mNGS results were more helpful in changed antibiotic therapy for patients with acute infections than those for patients with chronic

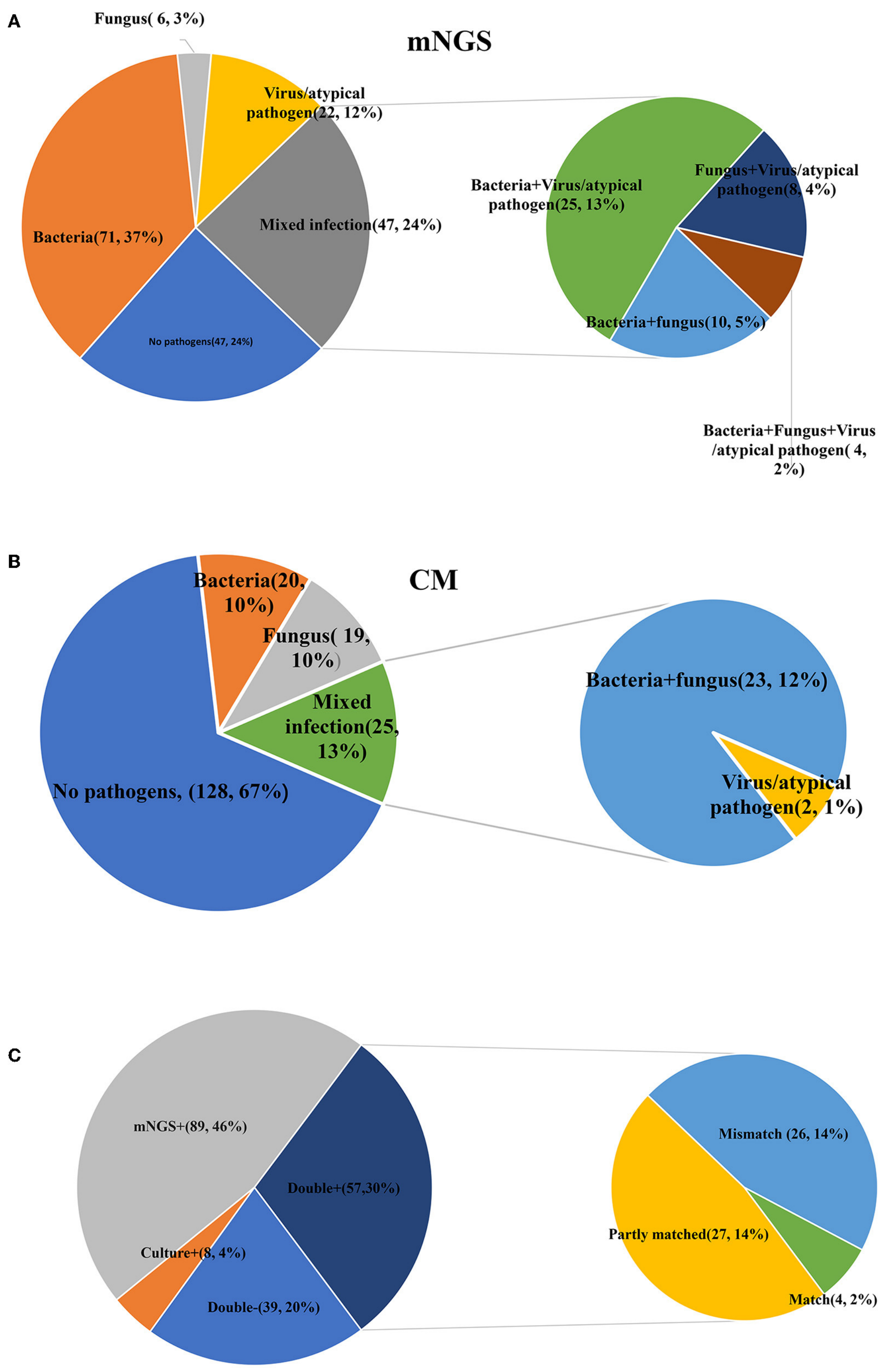


FIGURE 2
Comparisons between CM and mNGS in all patients. **(A)** The detection results of mNGS; **(B)** The detection results of CM; **(C)** The matched results between CM and mNGS. mNGS, metagenomic next-generation sequencing; CM, culture methods.

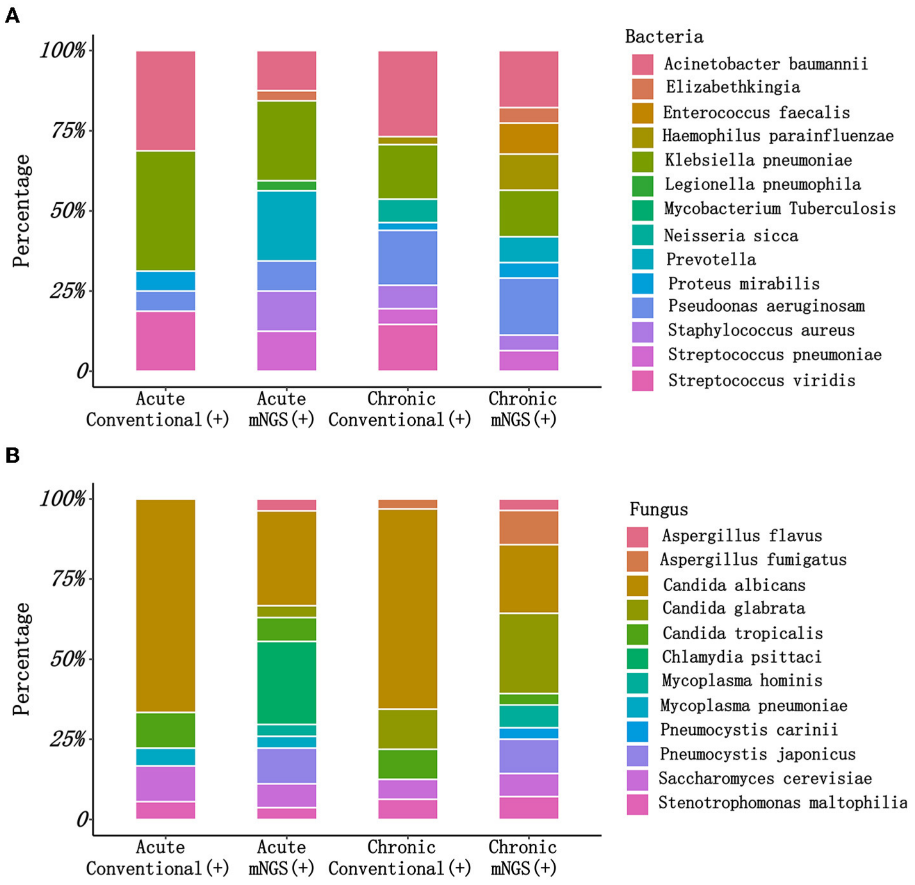


FIGURE 3
The pathogen types detected by CM and mNGS in different groups. **(A)** The bacteria types detected by CM and mNGS between acute infection group and chronic infection group; **(B)** The fungus types detected by CM and mNGS between acute infection group and chronic infection group. mNGS, metagenomic next-generation sequencing; CM, culture methods.

infections, with results suggesting more pronounced improvement of patients with acute infections. Finally, we found the related risk inflammatory factors of positive mNGS results in acute infection patients.

As a transformational and advanced technology, mNGS is able to detect a wide range of direct and potential infectious pathogens by sequencing of the extracted DNA from different specimens, and has the advantages of unbiased pathogen detection and short detection time (Finotello et al., 2018). Our study showed that the sensitivity, specificity, PPV and NPV were all higher than those of CM, especially the sensitivity, which is similar to some previously published research (Li et al., 2018; Cai et al., 2020). The high sensitivity of mNGS may be attributed to the long survival time of pathogenic genes in BALF, high detection rates of pathogens, and the small impact of antibiotic use on mNGS as compared with that in CM (Gosiewski et al., 2017). In addition, higher sensitivity contributed to higher detection rates of pathogens. Consistent with previous studies (Takeuchi et al., 2019; Ding et al., 2020; Huang et al., 2021), mNGS showed advantages in diagnosis of mixed infection of bacteria, fungi and viruses when compared to CM. Therefore, the relatively low matching rate between mNGS and CM may be attributed to the narrow detection range and low positive rate of CM. However, mNGS had more false-positive results than CM because mNGS could detect oral flora and colonizers in BALF from respiratory tract more easily. Sometimes, results of mNGS may blunt the diagnosis of pathogenicity resulting in the inability to distinguish among microbial infections, colonization and contaminations (Simner et al., 2018). Therefore, clinicians should interpret the mNGS results carefully when confronted with the inconsistent results of conventional methods after considering clinical manifestations and examination results of patients.

When it comes to the guidance of mNGS for clinical practice, we found that there were significantly more patients changed the antibiotic treatments in the positive group than negative group both in acute infection patients (60.5 vs. 28.0%, $P = 0.0022$) and chronic infection patients (46.2 vs. 22.6%, $P = 0.01112$). Consistently, a retrospective study enrolled 130 patients with acute respiratory failure mostly caused by pulmonary infection, and found that 58.5% of these patients had changed antibiotic regimen according to mNGS (Huang et al., 2021). For mNGS-positive patients where the CM was inconclusive, another study found that 58% of their patients were not covered by empirical antibiotics, and 61% of their patients modified antibiotic therapy based on mNGS (Miao et al., 2018). Some recent studies have demonstrated that mNGS has promising advantages in detecting antibiotic resistance genes (Ruppé et al., 2017; Chen et al., 2021a). Therefore, the result of

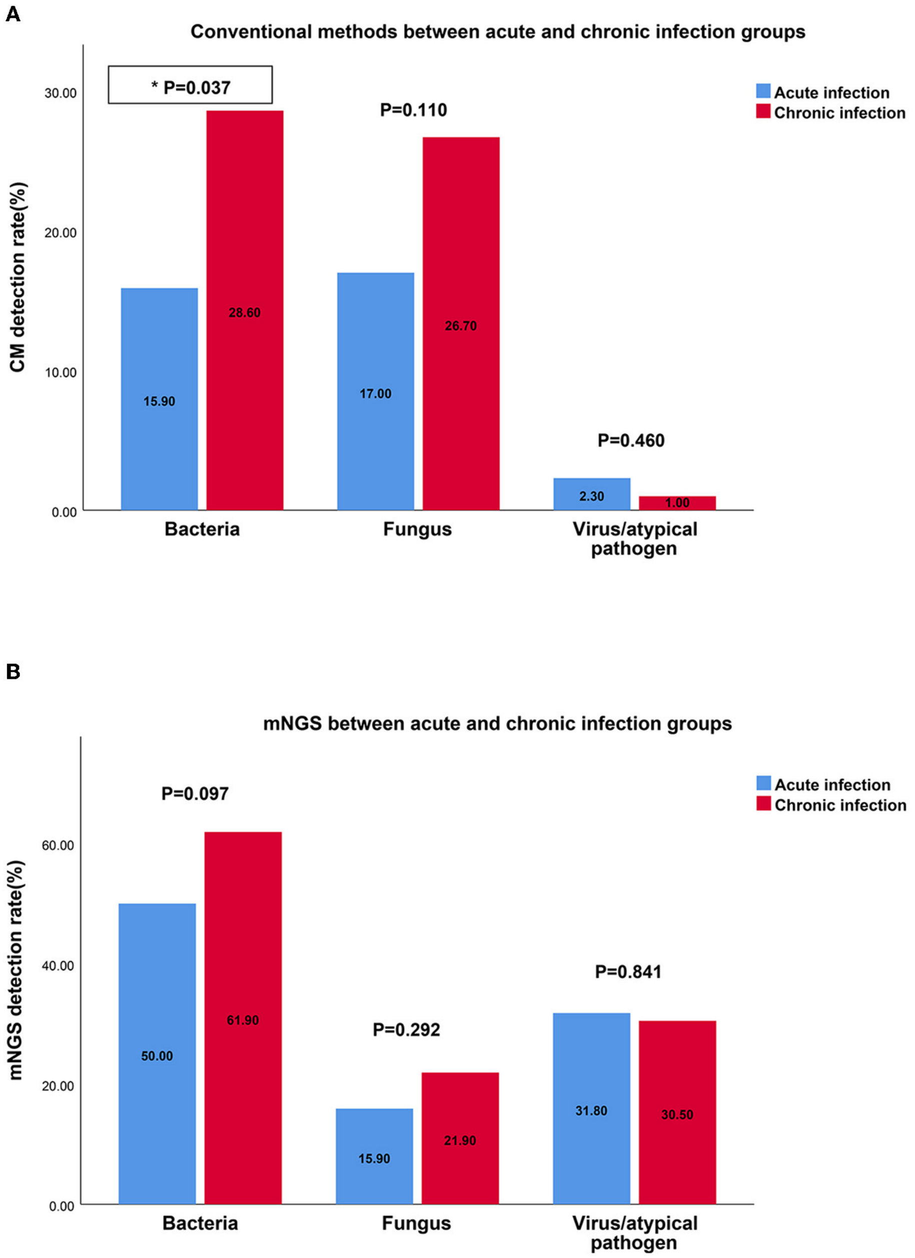


FIGURE 4
The detection rates of bacteria, fungus, virus by CM and mNGS in different groups. **(A)** The detection rate of bacteria, fungus, virus by CM between acute infection group and chronic infection group; **(B)** The detection rate of bacteria, fungus, virus by mNGS between acute infection group and chronic infection group. mNGS, metagenomic next-generation sequencing; CM, culture methods.

mNGS have been an essential reference to aid clinicians with disease diagnosis and targeted therapeutic schedule. However, there is no benefit for the survival prognosis after the change of antibiotics. In addition, survival analysis of our study showed no differences in the survival rate between positive group and negative groups whether in acute or chronic infection patients, and mNGS positive results had no significant impact on survival outcomes of these patients. A retrospective study enrolling 109 patients with infectious disease or not and collecting different samples including BALF reported that patients with positive mNGS results had higher 28-day mortality than those with negative mNGS results (9.0 vs. 0%, $P = 0.049$), but there was no significant difference in overall survival time, which is partly consistent with our view (Duan et al., 2021). But another retrospective study drew different conclusions that the average time of intensive care unit stay [β, −8.689 (95% CI, −16.176, −1.202); $P = 0.026$] and the time from onset to sequencing [β, −5.816 (95% CI, −9.936, −1.696); $P = 0.007$] of the mNGS-positive group were significantly shorter than those of the mNGS- negative group after analyzing 63 blood samples of critically ill patients, and more patients in mNGS-positive group changed the antibiotic treatment regimen after mNGS [OR, 3.789 (95% CI, 1.176, 12.211); $P < 0.001$] (Geng et al., 2021), which

TABLE 3 Comparisons between acute infection and chronic infection group.

Variables	Acute infection			Chronic infection		
	Positive (*n* = 38)	Negative (*n* = 50)	*P*	Positive (*n* = 52)	Negative (*n* = 53)	*P*
Hospital days (quartile)	9 (6.5, 12)	11 (8, 16)	0.831	10 (6.25, 17.5)	12 (8, 18)	0.311
Antibiotic			**0.0022**			**0.0112**
Changed	23 (60.5%)	14 (28.0%)		24 (46.2%)	12 (22.6%)	
Unchanged	15 (39.5%)	36 (72.0%)		28 (53.8%)	41 (77.4%)	
The decrease of CRP	*n* = 35	*n* = 25	0.766	*n* = 29	*n* = 24	0.626
Median (min, max)	56.8 (−87.0, 290)	33.3 (−54.0, 619)		25.5 (−44.3, 180)	40.0 (−128, 205)	
CRP return to normal	*n* = 35	*n* = 25	0.963	*n* = 29	*n* = 24	0.0959
Yes	11 (31.4%)	8 (32.0%)		5 (17.2%)	9 (37.5%)	
No	24 (68.6%)	17 (68.0%)		24 (82.8%)	15 (62.5%)	
The decrease of WBC	*n* = 15	*n* = 13	0.856	*n* = 16	*n* = 14	0.58
Median (min, max)	3.46 (−16.7, 12.0)	3.06 (−18.4, 13.5)		−0.328 (±5.27)	0.897 (±8.80)	
WBC return to normal	*n* = 15	*n* = 13	0.705	*n* = 16	*n* = 14	0.491
Yes	8 (53.3%)	6 (46.2%)		6 (37.5%)	7 (50.0%)	
No	7 (46.7%)	7 (53.8%)		10 (62.5%)	7 (50.0%)	
The decrease of neutrophil	*n* = 25	*n* = 18	0.331	*n* = 20	*n* = 17	0.351
Median (min, max)	12.4 (−6.30, 82.2)	16.3 (2.00, 54.7)		6.04 (±12.1)	1.92 (±14.1)	
Neutrophil return to normal	*n* = 25	*n* = 18	**0.0284**	*n* = 20	*n* = 17	0.969
Yes	16 (64.0%)	17 (94.4%)		6 (30.0%)	5 (29.4%)	
No	9 (36.0%)	1 (5.6%)		14 (70.0%)	12 (70.6%)	

PCT, procalcitonin; CRP, C-reactive protein; WBC, white blood cell; N, neutrophil. The bold values indicate that *p*-value is less than 0.05.

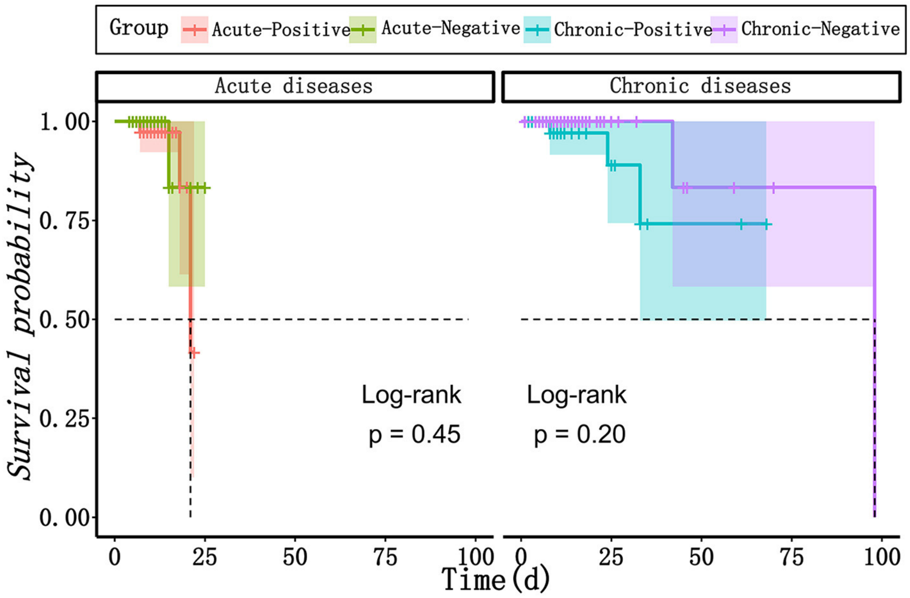
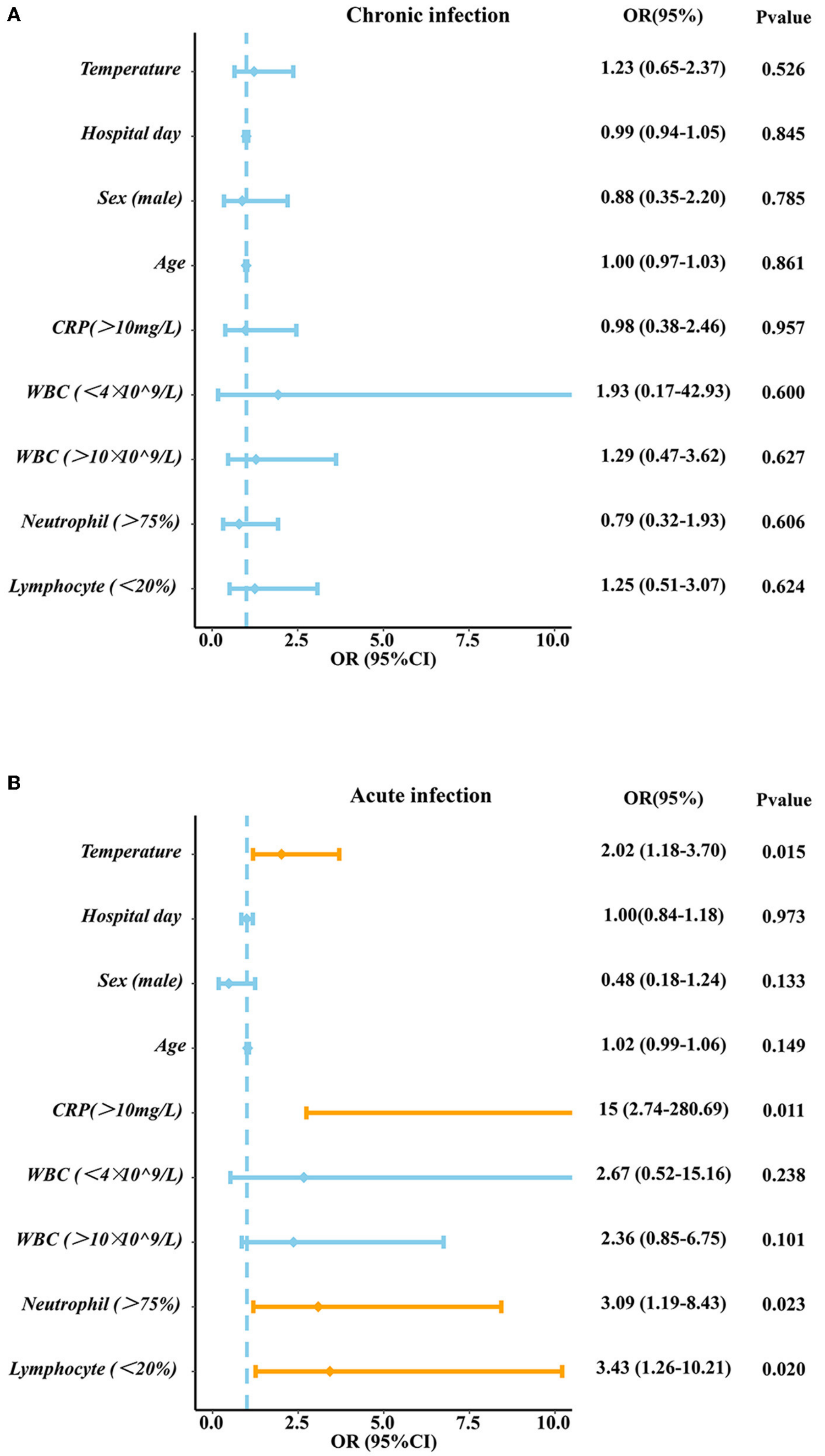


FIGURE 5
The survival curves between mNGS-positive group and mNGS negative group in patients with acute infection and chronic infection, respectively. mNGS, metagenomic next-generation sequencing.

TABLE 4 COX regression analysis between acute infection group and chronic infection group.

Variables	Acute infection group		Chronic infection group	
	HR (95%CI)	*P*	HR (95%CI)	*P*
Sex		0.372		0.695
Female	Reference		Reference	
Male	2.835 (0.288, 27.890)		0.621 (0.058, 6.700)	
Age		0.994		0.991
<45	Reference		Reference	
45–60	0 (0, 0)	0.980	6,962.448 (0, 7.210E+125)	
60–75	1.001 (0.090, 11.129)	0.999	1.026 (0, 1.039E+132)	
≥75	0.711 (0.043, 11.639)		97,260.235 (0, 1.001E+126)	
Anti-infective therapy		0.462		0.203
Unchanged	Reference		Reference	
Modified	2.353 (0.240, 23.046)		4.362 (0.451, 42.188)	
NGS		0.997		0.795
No pathogens	Reference		Reference	
Bacteria	0.237 (0.009, 6.332)	0.391	3,650.339 (0, 3.148E+47)	0.875
Fungus	0 (0, 0)	0.998	0.007 (0, 0)	0.998
Virus/atypical pathogen	0 (0, 0)	0.984	2,245.673 (0, 1.937E+47)	0.881
Bacteria + fungus	0.618 (0.025, 15.449)	0.769	0.008 (0, 0)	0.996
Bacteria + virus/atypical pathogen	0 (0, 0)	0.993	1.072 (0, 1.045E+62)	0.999
Fungus + virus/atypical pathogen	0.371 (0.011, 12.584)	0.582	33,627.634 (0, 2.906E+48)	0.840
Bacteria + fungus + virus/atypical	0 (0, 0)	0.997	0.008 (0, 0)	0.998
Pathogen				
mNGS		0.457		0.229
Negative	Reference		Reference	
Positive	2.409 (0.237, 24.488)		4.177 (0.407, 42.841)	
PCT		0.800		0.795
<0.5	Reference		Reference	
≥0.5	0.774 (0.107, 5.604)		0.039 (0, 1518968057)	
CRP		0.784		0.663
<10	Reference		Reference	
≥10	26.132 (0.000, 3.684E+11)		29.576 (0, 32876025.15)	
WBC		0.637		0.959
<10	Reference		Reference	
≥10	1.484 (0.287, 7.661)		0.938 (0.081, 10.848)	
N%		0.933		0.396
<75	Reference		Reference	
≥75	1.101 (0.117, 10.379)		58.127 (0.005, 692682.844)	
L%		0.638		0.535
<20	Reference		Reference	
≥20	1.517 (0.267, 8.606)		0.029 (0, 2040.882)	

mNGS, metagenomic next-generation sequencing; CM, culture methods; WBC, white blood cell; PCT, procalcitonin; CRP, C-reactiveprotein; N, neutrophil; L, lymphocyte; HR, hazard ratio; CI, confidence interval.



A Chronic infection

	OR(95%)	Pvalue
Temperature	1.23 (0.65-2.37)	0.526
Hospital day	0.99 (0.94-1.05)	0.845
Sex (male)	0.88 (0.35-2.20)	0.785
Age	1.00 (0.97-1.03)	0.861
CRP(>10mg/L)	0.98 (0.38-2.46)	0.957
WBC (<4×10^9/L)	1.93 (0.17-42.93)	0.600
WBC (>10×10^9/L)	1.29 (0.47-3.62)	0.627
Neutrophil (>75%)	0.79 (0.32-1.93)	0.606
Lymphocyte (<20%)	1.25 (0.51-3.07)	0.624

B Acute infection

	OR(95%)	Pvalue
Temperature	2.02 (1.18-3.70)	0.015
Hospital day	1.00(0.84-1.18)	0.973
Sex (male)	0.48 (0.18-1.24)	0.133
Age	1.02 (0.99-1.06)	0.149
CRP(>10mg/L)	15 (2.74-280.69)	0.011
WBC (<4×10^9/L)	2.67 (0.52-15.16)	0.238
WBC (>10×10^9/L)	2.36 (0.85-6.75)	0.101
Neutrophil (>75%)	3.09 (1.19-8.43)	0.023
Lymphocyte (<20%)	3.43 (1.26-10.21)	0.020

FIGURE 6
The Logistic analysis in acute infection group and chronic infection group. **(A)** The Logistic analysis in chronic infection group **(B)** The Logistic analysis in acute infection group. PCT, procalcitonin; CRP, C-reactive protein; WBC, white blood cell; N, neutrophil; L, lymphocyte.

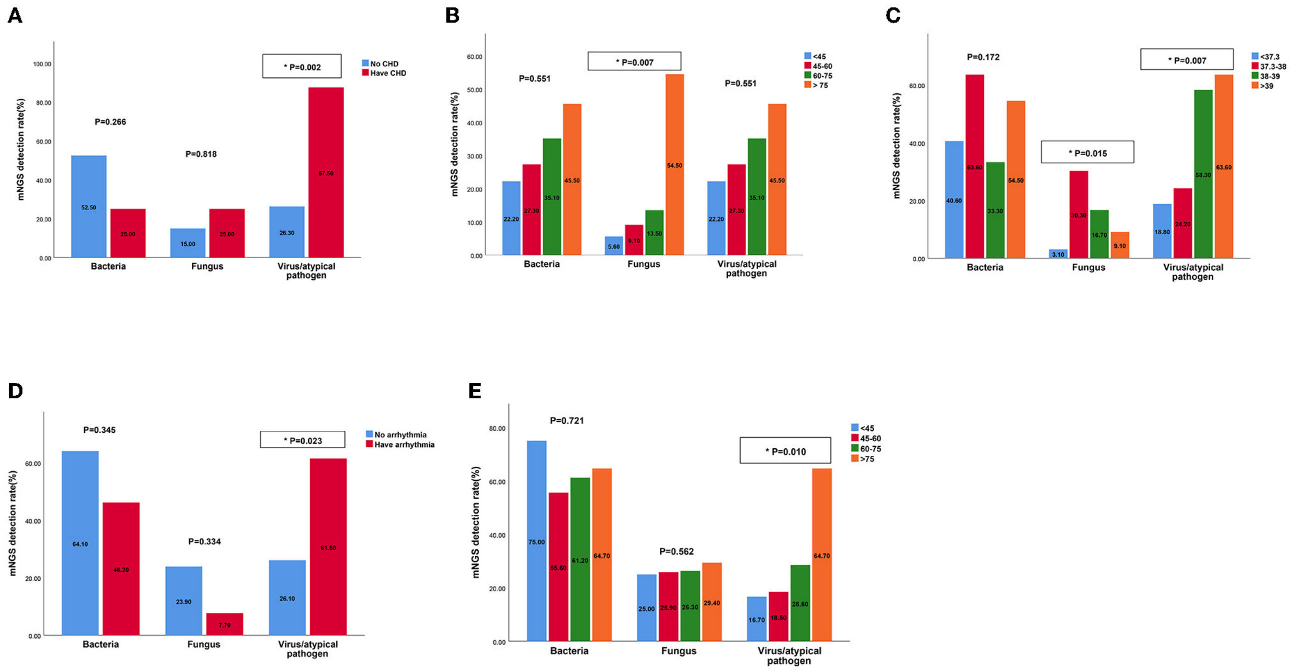


FIGURE 7
The detection rates of bacteria, fungus, virus by mNGS in patients with different characteristics of acute and chronic infection groups. **(A)** The detection rates of bacteria, fungus, virus by mNGS in patients with CHD in acute infection group; **(B)** The detection rates of bacteria, fungus, virus by mNGS in patients of different ages in acute infection group; **(C)** The detection rates of bacteria, fungus, virus by mNGS in patients with different temperatures in acute infection group; **(D)** The detection rates of bacteria, fungus, virus by mNGS in patients with arrhythmia in chronic infection group; **(E)** The detection rates of bacteria, fungus, virus by mNGS in patients of different ages in chronic infection group. mNGS, metagenomic next-generation sequencing; CHD, cornary heart diseases.

is consistent with the conclusion of another study enrolling 56 ICU patients with 131 samples including BALF (Liang et al., 2023). The reason for this inconsistency may be attributed to the differences of the severity of diseases. It was identified that mNGS was associated with a better diagnosis, treatments and prognosis of infectious patients, especially those critically ill patients with acute respiratory failure (Zhang et al., 2020; Xi et al., 2022). Therefore, mNGS could serve as a novel technology for infectious disease diagnosis and provide useful guidance on antibiotic strategies based on appropriate patient selection and scientific data analysis. Large-scale and prospective studies are required in future to verify the value and impact of mNGS-guided treatments in clinical practice.

Based on the advantages and clinical guidance of mNGS, we then investigated the related factors contributing to the positive mNGS results. It was found that high levels of temperature, CRP, neutrophil and low levels of lymphocyte may lead to the positive mNGS results in acute infection patients while no variable was identified to predict positive results in chronic infection patients. These laboratory indicators will help clinicians make decisions about the utilization of the mNGS. What's more, other characteristics of patients such as comorbidities and age were also identified to predict results of mNGS. Our study demonstrated that viral/atypical pathogens were more likely to be detected in acute and chronic infection patients complicated with heart diseases, which maybe because virus can replicate in cardiomyocyte, leading to heart dysfunction (Kenney et al., 2022). Differently, another study reported that APACHE II score (OR = 1.096), immune-related diseases (OR = 6.544), and hypertension (OR = 2.819) were considered as positive independent factors for mNGS positove results in patients with sepsis infection (Sun et al., 2022). We also found fungi were more likely to be detected in old aged patients in acute infection group, while viral/atypical pathogens were more likely to be detected in such patients of chronic infection group, probasbly because most elderly patients have immunity suppression, antibiotic resistance and comorbidity with other diseases (Aronen et al., 2019; Chen et al., 2021b). Similar to our study, a study used mNGS from plasma samples to dignose 40 travelers with acute fever (≥38°C), and found 11 patients were diagnosed with viral infection, highlighting the diagnostic value for acute high fever patients (Jerome et al., 2019).

Therefore, mNGS proved to be more valuable to acute infection patients than chronic infection patients. Various studies have proved the clinical value of mNGS for acute infectious patients. Some studies discussed the application of mNGS in acute viral encephalitis (Cao and Zhu, 2020), acute respiratory distress syndrome (Zhang et al., 2021; Wang et al., 2022), acute respiratory distress syndrome (Zhang et al., 2023) and other acute infectious diseases, demonstrating that mNGS is a promising tool for the diagnosis of acute disease caused by multiple infectious agents. mNGS represents a valuable supplementary tool to CM in order to rapidly determine etiological factors of various infections and guide treatment decision- making for patients (Xu et al., 2023). However, mNGS will interfere with the diagnosis of pathogenic bacteria when detecting broad-spectrum pathogens, resulting in an inability to distinguish among microbial infections, colonization and contaminations. The cost and quality surveillance of mNGS still need more efforts. Therefore, further studies are required to not only identify the complementary role of mNGS from different samples in order to support CM in routine clinical practice (Gu et al., 2021), but also explore more rapid, economic and targeted technology of mNGS to achieve early and accurate diagnosis of infectious diseases (Li et al., 2022).

In this study, we compared mNGS and CM in sensitivity, specificity, and pathogen types. On this basis, we also compared and analyzed the differences between the positive and negative groups of mNGS between acute and chronic infection patients in multicenters for the first time. Higher diagnostic value of metagenomic next- generation sequencing was demonstrated in patients with acute infection than patients with chronic infection in our study. However, the sample size in this study is relatively small. There is also a lack of randomized controls. And the retrospective nature of the study may miss some important data, which may bias the results and conclusion. Finally, the lack of a gold standard comparator for diagnostics, classification bias and antibiotic usage details may limit the generalizability of the conclusion of the present study.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the 10th People's Hospital affiliated to Tongji University and Tongren Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from a by-product of routine care or industry. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

AY: Formal analysis, Investigation, Methodology, Software, Supervision, Writing—original draft, Writing—review & editing. JW: Formal analysis, Methodology, Software, Writing—original draft. QX: Data curation, Formal analysis, Writing—original draft. BS: Data curation, Formal analysis, Writing—original draft. KS: Data curation, Writing—review & editing. FH: Data curation, Supervision, Validation, and Writing—review & editing. CW: Conceptualization, Data curation, Funding acquisition, Resources, Supervision, Validation, Writing—review & editing. SX: Conceptualization, Data curation, Funding acquisition, Resources, Supervision, Validation, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Shanghai Tongren Hospital-Donghua University School of Biology and Medical Engineering Medical Cross-Cooperation Project (2023DHYGJC-YBB05), and the Shanghai Chongming District Science and Technology Commission "Sustainable Development Science and Technology Innovation Action Plan" project (CKY2023-41).

Acknowledgments

The authors would like to thank all the patients who donate their biological samples.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1295184/full#supplementary-material

References

Aronen, M., Viikari, L., Kohonen, I., Vuorinen, T., Hämeenaho, M., Wuorela, M., et al. (2019). Respiratory tract virus infections in the elderly with pneumonia. *BMC Geriatr.* 19:111. doi: 10.1186/s12877-019-1125-z

Cai, Y., Fang, X., Chen, Y., Huang, Z., Zhang, C., Li, W., et al. (2020). Metagenomic next generation sequencing improves diagnosis of prosthetic joint infection by detecting the presence of bacteria in periprosthetic tissues. *Int. J. Infect. Dis.* 96, 573–578. doi: 10.1016/j.ijid.2020.05.125

Cao, J., and Zhu, X. Q. (2020). Acute viral encephalitis associated with human parvovirus B19 infection: unexpectedly diagnosed by metagenomic next-generation sequencing. *J. Neurovirol.* 26, 980–983. doi: 10.1007/s13365-020-00885-6

Chen, H., Bai, X., Gao, Y., Liu, W., Yao, X., Wang, J., et al. (2021a). Profile of bacteria with ARGs among real-world samples from ICU admission patients with pulmonary infection revealed by metagenomic. *NGS Infect. Drug Resist*. 14, 4993–5004. doi: 10.2147/IDR.S335864

Chen, L., Huang, H., and Chen, X. (2021b). Distribution of pathogens in elderly chinese patients with pneumonia: a systematic review and meta-analysis. *Front. Med.* 8:584066. doi: 10.3389/fmed.2021.584066

Ding, L., Zhao, Y., Li, X., Wang, R., Li, Y., Tang, X., et al. (2020). Early diagnosis and appropriate respiratory support for Mycoplasma pneumoniae pneumonia associated acute respiratory distress syndrome in young and adult patients: a case series from two centers. *BMC Infect. Dis.* 20:367. doi: 10.1186/s12879-020-05085-5

Duan, H., Li, X., Mei, A., Li, P., Liu, Y., Li, X., et al. (2021). The diagnostic value of metagenomic next? generation sequencing in infectious diseases. *BMC Infect. Dis.* 21:62. doi: 10.1186/s12879-020-05746-5

Finotello, F., Mastrorilli, E., Di Camillo, D., and Measuring, B. (2018). the diversity of the human microbiota with targeted next-generation sequencing. *Brief Bioinformat.* 19, 679–692.

GBD 2019 Chronic Respiratory Diseases Collaborators (2023). Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. *EClinicalMedicine.* 59:101936. doi: 10.1016/j.eclinm.2023.101936

Geng, S., Mei, Q., Zhu, C., Fang, X., Yang, T., Zhang, L., et al. (2021). Metagenomic next-generation sequencing technology for detection of pathogens in blood of critically ill patients. *Int. J. Infect. Dis.* 103, 81–87. doi: 10.1016/j.ijid.2020.11.166

Gosiewski, T., Ludwig-Galezowska, A. H., Huminska, K., Sroka-Oleksiak, A., Radkowski, P., Salamon, D., et al. (2017). Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia. *Eur. J. Clin. Microbiol. Infect. Dis.* 36, 329–336. doi: 10.1007/s10096-016-2805-7

Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., et al. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* 27, 115–124. doi: 10.1038/s41591-020-1105-z

Huang, C., Chen, H., Ding, Y., Ma, X., Zhu, H., Zhang, S., et al. (2021). A Microbial world: could metagenomic next-generation sequencing be involved in acute respiratory failure? *Front. Cell. Infect. Microbiol.* 11:738074. doi: 10.3389/fcimb.2021.738074

Jerome, H., Taylor, C., Sreenu, V. B., Klymenko, T., Filipe, A. D. S., Jackson, C., et al. (2019). Metagenomic next- generation sequencing aids the diagnosis of viral infections in febrile returning travellers. *J. Infect.* 79, 383–388. doi: 10.1016/j.jinf.2019.08.003

Jin, X., Li, J., Shao, M., Lv, X., Ji, N., Zhu, Y., et al. (2022). Improving suspected pulmonary infection diagnosis by bronchoalveolar lavage fluid metagenomic next-generation sequencing: a multicenter retrospective study. *Microbiol. Spectr.* 10:e0247321. doi: 10.1128/spectrum.02473-21

Kenney, A. D., Aron, S. L., Gilbert, C., Kumar, N., Chen, P., Eddy, A., et al. (2022). Influenza virus replication in cardiomyocytes drives heart dysfunction and fibrosis. *Sci. Adv.* 8:eabm5371. doi: 10.1126/sciadv.abm5371

Li, H., Gao, H., Meng, H., Wang, Q., Li, S., Chen, H., et al. (2018). Detection of pulmonary infectious pathogens from lung biopsy tissues by metagenomic next-generation sequencing. *Front. Cell. Infect. Microbiol.* 8:205. doi: 10.3389/fcimb.2018.00205

Li, S., Tong, J., Liu, Y., Shen, W., and Hu, P. (2022). Targeted next generation sequencing is comparable with metagenomic next generation sequencing in adults with pneumonia for pathogenic microorganism detection. *J. Infect.* 85, e127–e9. doi: 10.1016/j.jinf.2022.08.022

Li, Y., Sun, B., Tang, X., Liu, Y. L., He, H. Y., Li, X. Y., et al. (2020). Application of metagenomic next-generation sequencing for bronchoalveolar lavage diagnostics in critically ill patients. *Eur. J. Clin. Microbiol. Infect. Dis.* 39, 369–374. doi: 10.1007/s10096-019-03734-5

Liang, Y., Feng, Q., Wei, K., Hou, X., Song, X., Li, Y., et al. (2023). Potential of metagenomic next-generation sequencing in detecting infections of ICU patients. *Mol. Cell. Probes* 68:101898. doi: 10.1016/j.mcp.2023.101898

Lv, M., Zhu, C., Zhu, C., Yao, J., Xie, L., Zhang, C., et al. (2023). Clinical values of metagenomic next- generation sequencing in patients with severe pneumonia: a systematic review and meta-analysis. *Front. Cell. Infect. Microbiol.* 13:1106859. doi: 10.3389/fcimb.2023.1106859

Miao, Q., Ma, Y., Wang, Q., Pan, J., Zhang, Y., Jin, W., et al. (2018). Microbiological diagnostic performance of metagenomic next-generation sequencing when applied to clinical practice. *Clin. Infect. Dis.* 67(suppl_2):S231–S240. doi: 10.1093/cid/ciy693

Ruppé, E., Lazarevic, V., Girard, M., Mouton, W., Ferry, T., Laurent, F., et al. (2017). Clinical metagenomics of bone and joint infections: a proof of concept study. *Sci. Rep.* 7:7718. doi: 10.1038/s41598-017-07546-5

Simner, P. J., Miller, S., and Carroll, K. C. (2018). Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. *Clin. Infect. Dis.* 66, 778–788. doi: 10.1093/cid/cix881

Sun, L., Zhang, S., Yang, Z., Yang, F., Wang, Z., Li, H., et al. (2022). Clinical application and influencing factor analysis of metagenomic next-generation sequencing (mNGS) in ICU patients with sepsis. *Front. Cell. Infect. Microbiol.* 12:905132. doi: 10.3389/fcimb.2022.905132

Takeuchi, S., Kawada, J. I., Horiba, K., Okuno, Y., Okumura, T., Suzuki, T., et al. (2019). Metagenomic analysis using next-generation sequencing of pathogens in bronchoalveolar lavage fluid from pediatric patients with respiratory failure. *Sci. Rep.* 9:12909. doi: 10.1038/s41598-019-49372-x

Wang, J., Han, Y., and Feng, J. (2019). Metagenomic next-generation sequencing for mixed pulmonary infection diagnosis. *BMC Pulm. Med.* 19:252. doi: 10.1186/s12890-019-1022-4

Wang, R., Feng, R., Xia, C., Ruan, F., Luo, P., Guo, J., et al. (2022). Early detection of gram-negative bacteria using metagenomic next-generation sequencing in acute respiratory distress syndrome: a case report. *Exp. Ther. Med.* 24:573. doi: 10.3892/etm.2022.11510

Xi, Y., Zhou, J., Lin, Z., Liang, W., Yang, C., Liu, D., et al. (2022). Patients with infectious diseases undergoing mechanical ventilation in the intensive care unit have better prognosis after receiving metagenomic next-generation sequencing assay. *Int. J. Infect. Dis.* 122, 959–969. doi: 10.1016/j.ijid.2022.07.062

Xiao, Y. H., Liu, M. F., Wu, H., Xu, D. R., and Zhao, R. (2023). Clinical efficacy and diagnostic value of metagenomic next-generation sequencing for pathogen detection in patients with suspected infectious diseases: a retrospective study from a large tertiary hospital. *Infect. Drug Resist.* 16, 1815–1828. doi: 10.2147/IDR.S401707

Xu, J., Zhou, P., Liu, J., Zhao, L., Fu, H., Han, Q., et al. (2023). Utilizing metagenomic next-generation sequencing (mNGS) for rapid pathogen identification and to inform clinical decision-making: results from a large real-world cohort. *Infect. Dis. Therapy* 12, 1175–1187. doi: 10.1007/s40121-023-00790-5

Zhang, B., Chen, X., Yao, X., Li, M., Li, Z., Liu, B., et al. (2023). The diagnostic value of blood metagenomic next-generation sequencing in patients with acute hematogenous osteomyelitis. *Front. Cell. Infect. Microbiol.* 13:1106097. doi: 10.3389/fcimb.2023.1106097

Zhang, P., Chen, Y., Li, S., Li, C., Zhang, S., Zheng, W., et al. (2020). Metagenomic next-generation sequencing for the clinical diagnosis and prognosis of acute respiratory distress syndrome caused by severe pneumonia: a retrospective study. *PeerJ* 8:e9623. doi: 10.7717/peerj.9623

Zhang, X. J., Zheng, J. Y., Li, X., Liang, Y. J., and Zhang, Z. D. (2021). Usefulness of metagenomic next-generation sequencing in adenovirus 7-induced acute respiratory distress syndrome: a case report. *World J. Clin. Cases* 9, 6067–6072. doi: 10.12998/wjcc.v9.i21.6067

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 02 February 2024
DOI 10.3389/fmicb.2023.1272636

Genomic analysis of *Ralstonia pickettii* reveals the genetic features for potential pathogenicity and adaptive evolution in drinking water

Chao Yuan[1,2,3,4]*, Tianfeng An[1,2,3,4], Xinlong Li[4], Jiao Zou[4], Zhan Lin[4], Jiale Gu[4], Ruixia Hu[2,3,4] and Zhongze Fang[1,2,3,4]*

[1]Department of Toxicology and Sanitary Chemistry, School of Public Health, Tianjin Medical University, Tianjin, China, [2]Tianjin Key Laboratory of Environment, Nutrition and Public Health, Tianjin Medical University, Tianjin, China, [3]Center for International Collaborative Research on Environment, Nutrition and Public Health, School of Public Health, Tianjin Medical University, Tianjin, China, [4]School of Public Health, Tianjin Medical University, Tianjin, China

Check for updates

OPEN ACCESS

EDITED BY
Tulio de Lima Campos,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Munusamy Madhaiyan,
Agency for Science, Technology and Research, Singapore
Michael P. Ryan,
University of Limerick, Ireland

*CORRESPONDENCE
Chao Yuan
✉ hellgel_yc@126.com
Zhongze Fang
✉ fangzhongze_temp@126.com

RECEIVED 04 August 2023
ACCEPTED 13 December 2023
PUBLISHED 02 February 2024

CITATION
Yuan C, An T, Li X, Zou J, Lin Z, Gu J, Hu R and Fang Z (2024) Genomic analysis of *Ralstonia pickettii* reveals the genetic features for potential pathogenicity and adaptive evolution in drinking water.
Front. Microbiol. 14:1272636.
doi: 10.3389/fmicb.2023.1272636

Ralstonia pickettii, the most critical clinical pathogen of the genus *Ralstonia*, has been identified as a causative agent of numerous harmful infections. Additionally, *Ralstonia pickettii* demonstrates adaptability to extreme environmental conditions, such as those found in drinking water. In this study, we conducted a comprehensive genomic analysis to investigate the genomic characteristics related to potential pathogenicity and adaptive evolution in drinking water environments of *Ralstonia pickettii*. Through phylogenetic analysis and population genetic analysis, we divided *Ralstonia pickettii* into five Groups, two of which were associated with drinking water environments. The open pan-genome with a large and flexible gene repertoire indicated a high genetic plasticity. Significant differences in functional enrichment were observed between the core- and pan-genome of different groups. Diverse mobile genetic elements (MGEs), extensive genomic rearrangements, and horizontal gene transfer (HGT) events played a crucial role in generating genetic diversity. In drinking water environments, *Ralstonia pickettii* exhibited strong adaptability, and the acquisition of specific adaptive genes was potentially facilitated by genomic islands (GIs) and HGT. Furthermore, environmental pressures drove the adaptive evolution of *Ralstonia pickettii*, leading to the accumulation of unique mutations in key genes. These mutations may have a significant impact on various physiological functions, particularly carbon metabolism and energy metabolism. The presence of virulence-related elements associated with macromolecular secretion systems, virulence factors, and antimicrobial resistance indicated the potential pathogenicity of *Ralstonia pickettii*, making it capable of causing multiple nosocomial infections. This study provides comprehensive insights into the potential pathogenicity and adaptive evolution of *Ralstonia pickettii* in drinking water environments from a genomic perspective.

KEYWORDS

***Ralstonia pickettii*, pan-genome, core-genome, pathogenicity, adaptive evolution**

Introduction

The genus *Ralstonia*, a group of aerobic gram-negative bacteria belonging to the family Burkholderiaceae, was first described in 1995 (Yabuuchi et al., 1995). Currently, seven species have been identified: *Ralstonia insidiosa*, *Ralstonia mannitolilytica*, *Ralstonia pickettii*, *Ralstonia pseudosolanacearum*, *Ralstonia solanacearum*, and *Ralstonia syzygii*, and *Ralstonia wenshanensis* (Euzéby, 1991). *Ralstonia* species have increasingly been recognized as emerging nosocomial pathogens, particularly in immunocompromised patients (Ryan et al., 2006). Among the *Ralstonia* species, *Ralstonia pickettii*, previously known as *Burkholderia pickettii*, is considered the most clinically significant pathogen. While it is generally associated with low pathogenicity, occasional nosocomial outbreaks of *Ralstonia pickettii* infections have been reported (Nasir et al., 2019; Menekşe et al., 2022). *Ralstonia pickettii* could be isolated from various clinical specimens, including blood, wounds, urine, ear, nose swabs, or cerebrospinal fluid (Stelzmueller et al., 2006). Despite its low virulence, it has been implicated in causing potentially harmful infections and even death (Chen et al., 2015). However, there is limited research on the genetic diversity and potential pathogenicity of *Ralstonia pickettii* (Ryan et al., 2011).

On the other hand, the presence of *Ralstonia pickettii* in mineral solutions such as sterile saline solution, disinfectant, or other medical solutions, specifically purified water supplies, could explain its association with nosocomial outbreaks (Riley and Weaver, 1975; Ryan and Adley, 2013; Baker et al., 2022). Interestingly, *Ralstonia pickettii* has even been isolated from the potable water system of the International Space Station (PRJNA493516 from NCBI database). This highlights the adaptability of *Ralstonia pickettii* to survive in diverse environments, including drinking water, which poses a potential public health threat. Drinking water is a unique environment characterized by limited organic nutrient availability but suitable mineral concentrations for biological requirements (Liu et al., 2021). It is not an ideal condition for the survival of heterotrophic bacteria. However, *Ralstonia pickettii* has demonstrated remarkable environmental adaptability and the ability to thrive under extreme conditions, such as drinking water. This adaptability raises concerns about its potential to cause infectious diseases in future. The genetic mechanisms underlying its adaptation to drinking water environments are currently unknown.

Although *Ralstonia pickettii* is not considered a primary pathogen and is generally believed to have low virulence, its wide distribution in human-related environments and its potential to cause harmful infections emphasize the need for appropriate medical strategies and further in-depth research. Whole-genome sequencing (WGS) provides a valuable tool for exploring evolutionary relationships, genomic characteristics, and biotechnological properties at the gene level (Yuan et al., 2019; Yin et al., 2020). Genome plasticity enables bacteria to rapidly evolve and adapt to environmental variations by modulating virulence and antimicrobial resistance, in addition to environmental adaptation (Everitt et al., 2014; Yuan et al., 2020). The selective pressure on the core genome indicates adaptive evolution under specific environmental conditions. In this study, we present a comprehensive genomic analysis to investigate the genetic diversity, potential pathogenicity, and adaptive evolution of *Ralstonia pickettii* in drinking water.

Materials and methods

Genome data collection

All collected genomes of *Ralstonia pickettii* species were downloaded from the NCBI GenBank database. The genome completeness and contamination assessment were performed using CheckM v1.0.13 (Parks et al., 2015). A total of 76 available WGS data of *Ralstonia pickettii* were collected from NCBI and analyzed in this research. The information of these strains is presented in Supplementary Table S1.

Phylogenetic analysis based on single-copy core-gene orthologous gene families

The orthologous groups of protein families of pan-genome were delimited using OrthoFinder v2.5.1 (Emms and Kelly, 2015) software, employing the DIAMOND algorithm (Buchfink et al., 2015) with the default parameters. Subsequently, the single-copy orthologous gene families, core-gene families, accessory-gene families, and pan-gene families were extracted from the output results of OrthoFinder. Nucleotide sequences of the single-copy orthologous gene families were extracted and then aligned using MAFFT v7.475 (Katoh and Standley, 2013). For the phylogenetic analysis of *Ralstonia pickettii*, single-nucleotide polymorphisms (SNPs) present in the single-copy orthologous gene families were utilized. The Maximum Likelihood (ML) tree was constructed using MEGA 11 software (Kumar et al., 2016) [with the General Time Reversible (GTR) model].

Core- and pan-genome statistics analysis

Heap's law was applied to analyze the pan-genome models in this study. The relationship between the total number of gene families (*n*, y-axis) and the increasing number of genomes (*N*, x-axis) was examined. The curve fitting with a power-law regression is based on Heaps' law ($n = : N^{\gamma}$),where N represents the number of genomes, κ is a proportionality constant, and the growth exponent $\gamma > 0$ indicates an open pan-genome. This regression analysis provides insights into the expansion of the pan-genome as more genomes are included (Heaps, 1978; Tettelin et al., 2008). Additionally, the core-genome analysis was performed using regression analysis, and curve fitting of core-genome was performed using an exponential regression model ($n = \kappa \,\text{exp.}\, (mN) + \Theta$), where N represents the number of genomes and κ and m are proportionality constants (Bottacini et al., 2010). Descriptive statistical analysis was generated using OriginPro 9 software with exponential function model YldFert1 (core-genome) and power function model Allometric1 (pan-genome).

Genetic population structure analysis and gene functional category

The average nucleotide identity (ANI) was calculated using JSpecies v1.2.1 software (Richter and Rosselló-Móra, 2009). Population structure analysis was conducted using RhierBAPS v1.1.3 (Cheng et al., 2013; Tonkin-Hill et al., 2018). We analyzed the functional category of the gene family based on the Cluster of Orthologous Group (COG) assignment (Galperin et al., 2015). The functional annotation of proteins was performed using eggNOG-mapper v2.1.9 with default parameters (Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). The prophages were identified using the Phage Search Tool Enhanced Release (PHASTER; Arndt et al., 2016). Genomic islands were predicted using the IslandViewer 4 database with default parameters (Bertelli et al., 2017). HGTector (Zhu et al., 2014) was employed to identify the potential horizontal genes in *Ralstonia pickettii* species with default parameters.

Identification of virulence genes and resistance genes

To identify the virulence genes and resistance genes, protein sequences of all *Ralstonia pickettii* genomes were aligned using BLASTp against the data set from the Pathogen Host Interactions database (PHI-base 5.0; Urban et al., 2020) and Comprehensive Antibiotic Database (CARD; Jia et al., 2017) with three screening thresholds: (1) the percentage of identical < 50% or coverage < 60%; (2) the percentage of identical >50% and percentage of identical <75% and coverage > 60%; (3) the percentage of identical > 75% and coverage > 60%. All blast results in e-value cutoff of $<1e-6$. These results were visualized using the R packages pheatmap v1.0.12 and Adobe Illustrator CS6.

Identification of macromolecular systems

The detection and visualization of Macromolecular systems in *Ralstonia pickettii* species were performed using MacSyFinder (Abby et al., 2014) and TXSScan (Abby and Rocha, 2017) within Galaxy workflow[1] with the default parameters. The T4SS and T6SS were further analyzed using SecReT4 (Liu et al., 2012) and SecReT6 (Li et al., 2015) on the default parameters, respectively.

Comparative core-genomic and pan-genomic analysis

The comparative core-genomic and pan-genomic analysis was conducted to investigate the genetic characteristics among different groups. Increased and decreased core-gene families were extracted and annotated by the KEGG database (database resolution through Python 3.8, based on public information from the KEGG database). Two different classification methods counted results of “Gene function” and “Pathway” and shown in a bar chart. Based on gene data above, Relative Enrichment Ratio (RER) was introduced to describe the correlation between gene enrichment and pathway (Xu and Yuan, 2022). The RER was calculated using the formula: RER = (number of genes annotated in this pathway)/(total number of genes involved in this pathway) and shown in a line chart; the RER results were screened by 0.1 and 0.2 as the threshold (maximum number of peaks can be retained), and the pathway corresponding to each peak was extracted. All figures were made by GraphPad prism 9.0, R packages, and Adobe Illustrator CS6.

Results

Phylogenetic and genetic population analyses of *Ralstonia pickettii*

A maximum-likelihood phylogenetic tree was constructed using 3,010 concatenated single-copy orthologous gene families from 76 *Ralstonia pickettii* genomes (Figure 1). To explore the genomic similarities among the strains, genetic population structure analysis was performed using Bayesian analysis of population structure (BAPS) at two levels. This analysis categorized the *Ralstonia pickettii* strains into 5 BAPS classes at Level 1 or 10 BAPS classes at Level 2. The results from the BAPS analysis were consistent with the genetic distances estimated by the average nucleotide identity (ANI) values (Figure 1). Combining the results from the phylogenetic analysis, BAPS analysis, and ANI values, the *Ralstonia pickettii* strains were divided into five groups. The majority of the strains (64/76, 84.2%) were isolated from water-related environments. Specifically, 61 strains, mainly distributed in the Group 2 and Group 5, were isolated from the potable water system in the International Space Station, with isolation dates ranging from 2009 to 2015. Five strains (5/76, 6.6%, mainly distributed in Group 3 and Group 4) were isolated from humans (*Homo sapiens*), suggesting potential pathogenicity. These findings indicate that two clusters of strains from the International Space Station may represent different initial contaminants from Earth, rather than two evolutionary paths of a single contaminant. However, both clusters were exposed to the same environmental conditions, specifically drinking water.

Pan-genomic analysis revealed the genetic characteristics of *Ralstonia pickettii*

To further characterize the genetic characteristics of *Ralstonia pickettii*, pan-genome analysis was performed. A total of 10,005 pan-genome gene families were identified (Figure 2B). Among these, 3,514 (35.1%) represented the core-genome, and the remaining 6,491 (54.9%) represented the accessory genome (4,995, 49.9%) and strain-specific genes (1,496, 15.0%). Cluster of orthologous group (COG) annotation analysis was performed to categorize the function of pan-gene families. As shown in Figure 2C, the core-genome was significantly enriched in categories, such as COG-J (translation, ribosomal structure, and biogenesis), COG-E (amino acid transport and metabolism), COG-F (nucleotide transport and metabolism), COG-H (coenzyme transport and metabolism), and COG-I (lipid transport and metabolism) [COG-J, -E, and -I: Fisher’s exact test p-value < 0.05; COG-F, and -H: Fisher’s exact test p-value < 0.01]. These genes are primarily involved in maintaining normal physiological functions and material metabolism in bacteria. On the other hand, the accessory genes were enriched in categories, such as COG-L (replication, recombination, and repair), COG-N (cell motility), and COG-U (intracellular trafficking，secretion, and vesicular transport). These findings suggest that movement in different

[1] https://galaxy.pasteur.fr/

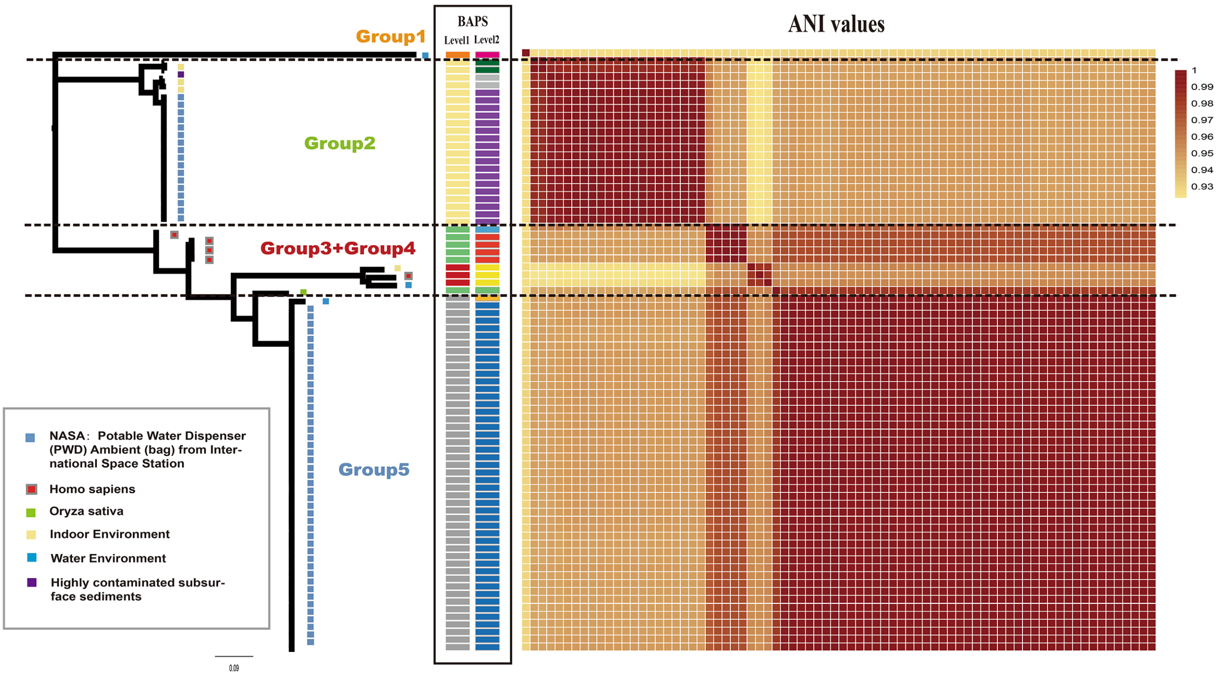


FIGURE 1
Phylogenetic relationship and isolated information of *Ralstonia pickettii*. The maximum likelihood (ML) tree was constructed using SNPs identified across 3,010 single-copy core gene families that are shared among the 76 *Ralstonia pickettii* genomes. On the right and middle of the chart, a heatmap representing the average nucleotide identities (ANI) and BAPS (two levels) is displayed. The isolated information of *Ralstonia pickettii* strains is labeled with different colors within gray frames.

environments and interactions with various materials are necessary for the development of new traits.

The pan-genome accumulation curve, representing the increasing number of genomes, followed Heaps' law ($n = \ :N^{\gamma}$) pan-genome model (Figure 2A), with $\gamma = 0.1423$. A positive exponent ($\gamma > 0$) indicated an open pan-genome, suggesting that novel accessory gene families that may be identified as additional strains are sampled. The pan-genome of Group 2, Group 3 + Group 4, and Group 5 also exhibited a linear upward trend (the strain number of Group 1 was too small and will not be discussed in this section). However, different power law values suggested different degrees of openness within each group. As shown in Figure 2A, the pan-genome of Group 3 + Group 4 had a higher exponent ($\gamma = 0.2078$), indicating a larger source gene pool and potential for adapting to new niches by acquiring novel genetic elements. In contrast, Group 2 ($\gamma = 0.0950$) and Group 5 ($\gamma = 0.0258$) experienced relatively simpler environmental pressures during evolution. The results of the pan-genome all-blast-all analysis are shown in Figure 2D (detailed information of pan-blast-screen results is shown in Supplementary Table S2), which is consistent with our population genetic analysis.

Genetic plasticity and genomic evolution mediated by numerous MGEs and HGT

Mobile genetic elements (MGEs) and horizontal gene transfer (HGT) play crucial roles in genetic diversity and the expansion of gene pools of bacteria (Ochman et al., 2000; Gyles and Boerlin, 2014). In the case of *Ralstonia pickettii*, we analyzed the distribution of MGEs and HGT events (Figure 3A; Supplementary Table S3). On average, each genome contained 17.3 ± 8 genomic islands (GIs), 33.4 ± 1 prophages, and 1078.3 ± 34.5 HGT genes. There were significant differences in the number and gene content of prophages (prophage genes) among the different groups. Group 5 had the highest number of prophages (20.2 ± 2.8), while Group 2 had a lower number of prophages (13.1 ± 0.5). However, Group 5 had a lower gene content of prophages compared with the other groups. Similar variations were observed in the number and gene content of GIs (GI genes) and HGT genes (HGT genes) among the groups. The distribution of MGEs and HGT genes in *Ralstonia pickettii* contributes to the formation of genomic diversity during evolution. After COG annotation analysis (Figure 3B), we observed that Group 2 and Group 5 had relatively fewer prophage genes in almost every COG category. However, GI genes of Group 2 and Group 5 were significantly enriched in COG-B (chromatin structure and dynamics) and COG-M (cell wall/membrane/envelope biogenesis), while HGT genes of Group 2 and Group 5 were significantly enriched in COG-L (replication, recombination, and repair). These specific gene distributions are likely related to the environmental adaptation of *Ralstonia pickettii* in the drinking water of the International Space Station, where strains from Group 2 and Group 5 are predominantly found. Further investigation revealed that strains isolated from drinking water had a higher number of prophages but a smaller number of prophage genes compared with the other strains (Figure 3C). This indicates differences in prophage species. More HGT genes were observed in strains isolated from drinking water, while no significant differences were found in the number of GIs and GI genes. Consistent with the previous results, GI genes were significantly enriched in COG-B and COG-M, while HGT genes were

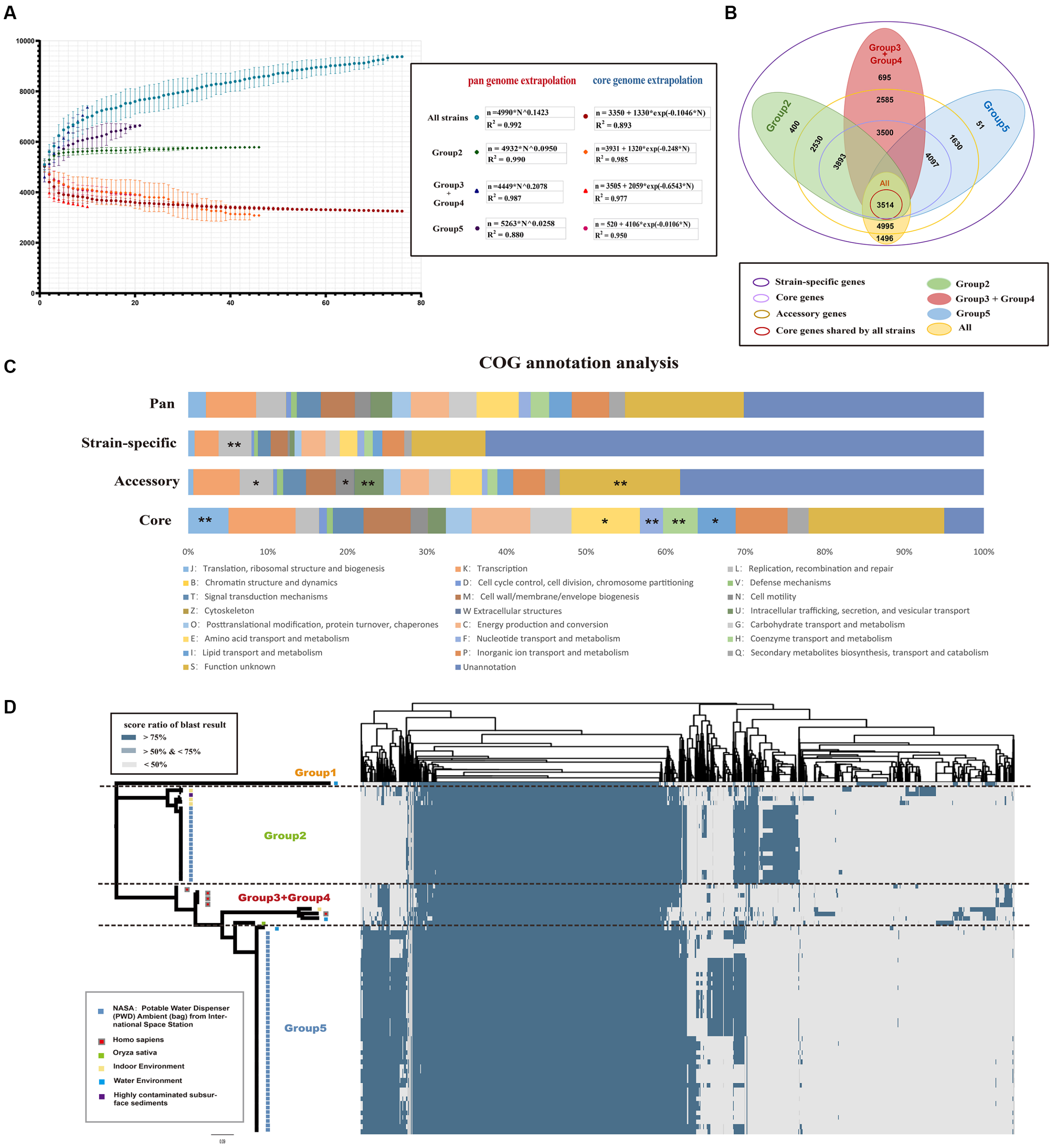


FIGURE 2
Core and pan-genomic analysis of *Ralstonia pickettii*. **(A)** Progressive curves were estimated for the core genome and pan-genome of all strains, as well as for different groups (Group 2, Group 3 + Group 4, and Group 5). These curves demonstrate a decrease in the number of core gene families and an increase in the number of pan-gene families as more genomes are added. Mathematical functions describing the pan-genome curves are displayed within the frame. **(B)** A Venn plot illustrates the gene content, including core genome, accessory genome, and strain-specific genes, among different groups of *Ralstonia pickettii* genomes. **(C)** The distribution of cluster of orthologous group (COG) categories is shown for the pan-genome, core genome, accessory genome, and strain-specific genes. Statistical significance is denoted by asterisks (*Fisher's exact test p-value < 0.05; **Fisher's exact test p-value < 0.01). **(D)** A cluster heatmap presents the pan-genome of *Ralstonia pickettii*, showing the clustering patterns of gene families.

significantly enriched in COG-L in *Ralstonia pickettii* isolated from drinking water. The diverse distribution of MGEs and HGT genes drive genetic plasticity and genomic evolution in *Ralstonia pickettii*, particularly in complex genetic backgrounds. Through these processes, *Ralstonia pickettii* can acquire novel metabolic properties to adapt to extreme environmental conditions.

Comparative core-genomic and pan-genomic analysis between different groups

The genetic characteristics observed in different groups of *Ralstonia pickettii* can be attributed to the development of unique

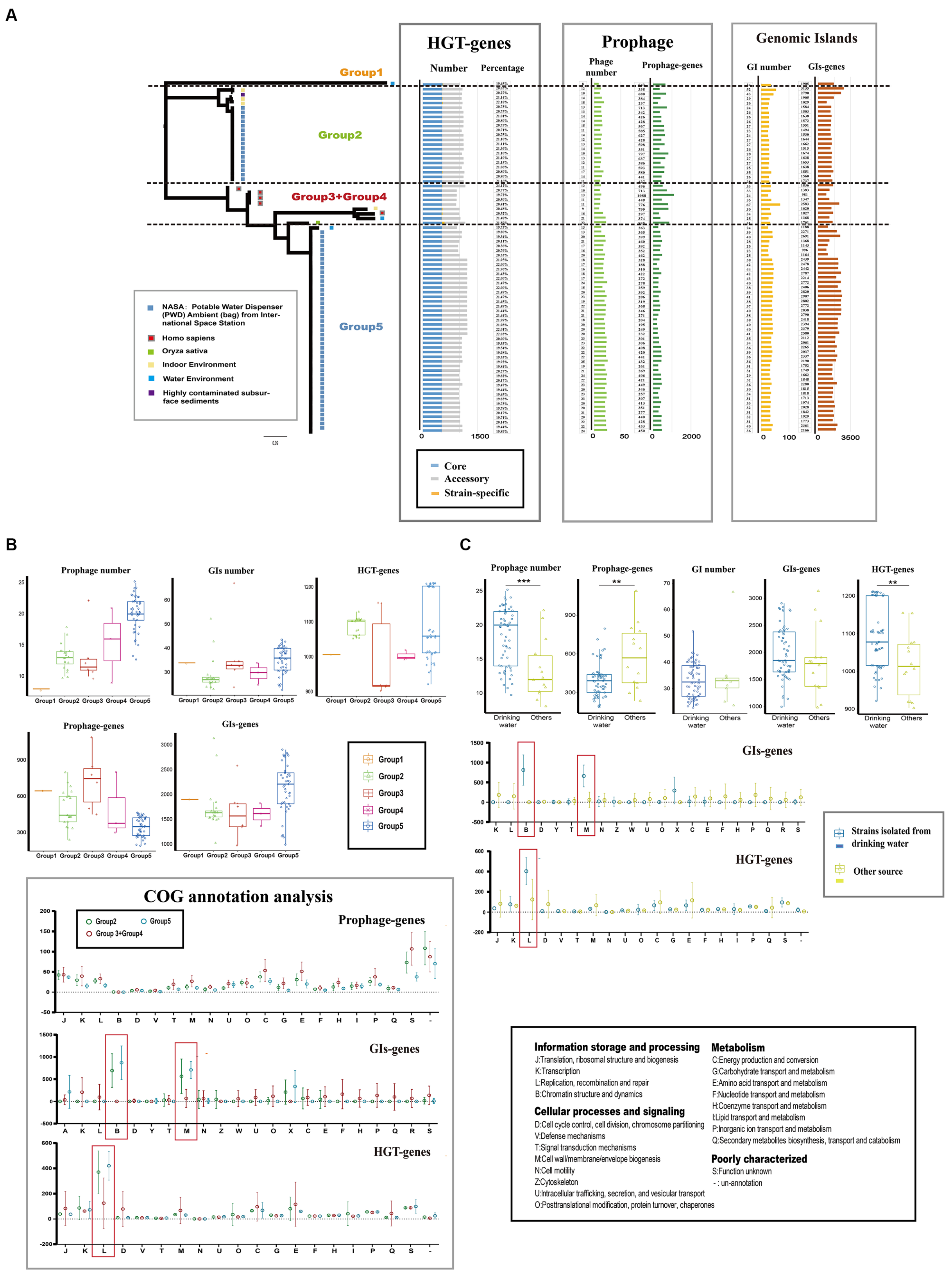
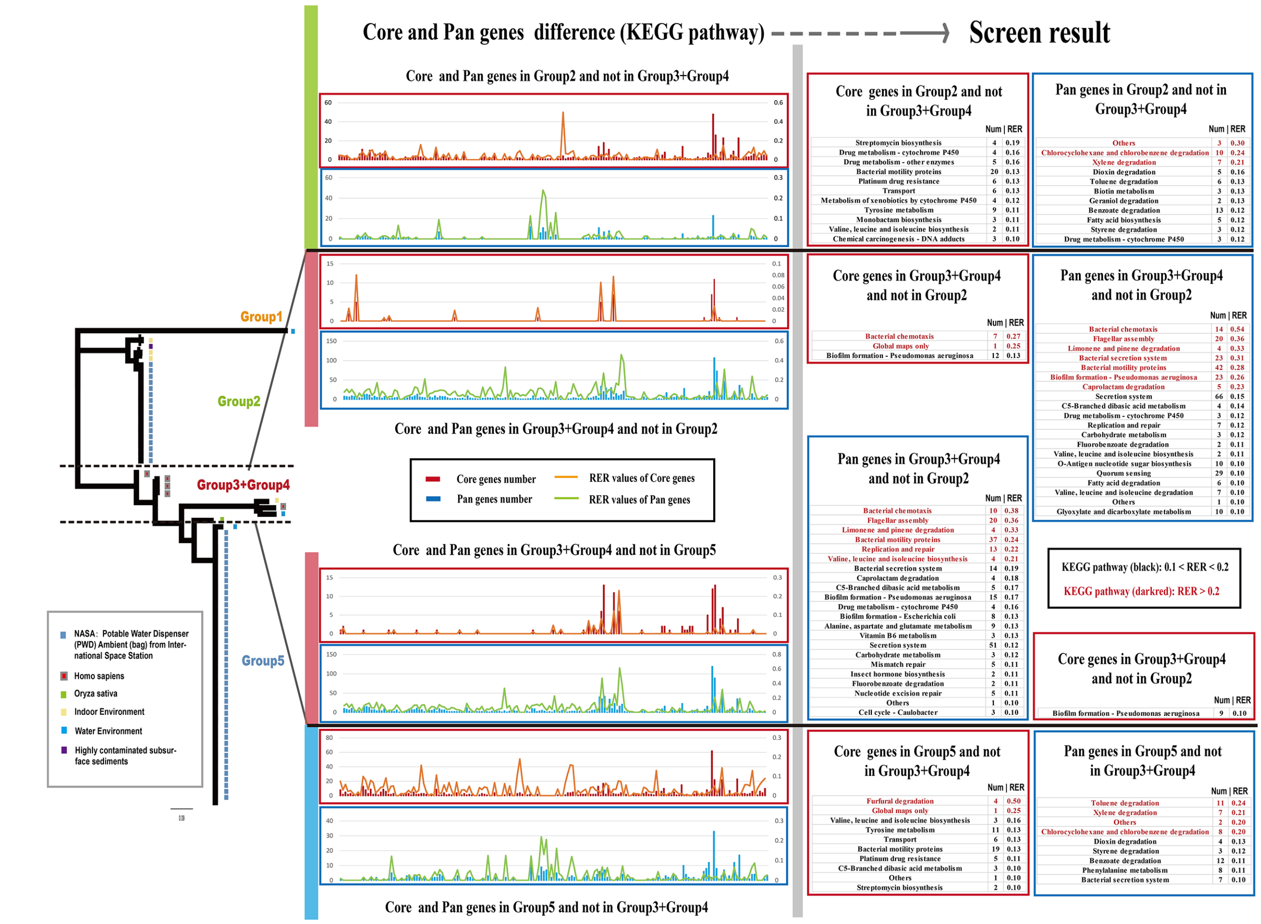


FIGURE 3
Mobile genetic elements (MGEs) and Horizontal gene transfers (HGTs) events in *Ralstonia pickettii*. **(A)** The distribution of mobile genetic elements (MGEs) and horizontally acquired genes in each strain is depicted. **(B)** The gene quantity and functional analysis of MGEs and horizontal gene transfer (HGT) events in different groups are presented. **(C)** A comparison is made between MGEs and horizontally acquired genes in strains isolated from drinking water and those obtained from other sources. Functional categories that show significant enrichment are highlighted with red boxes and were subjected to statistical testing (*t*-test). Statistical significance is indicated by asterisks (***t*-test *p*-value < 0.01; ****t*-test *p*-value < 0.001).

core-genomes and pan-genomes during the adaptive evolutionary process under different genetic backgrounds. To understand the genomic characteristics of each group, a KEGG function enrichment analysis was conducted on the core-genomes and pan-genomes of each group (Supplementary Figure S1; Supplementary Table S9). The results were presented using two methods: gene function and pathway. However, no significant differences were found between the core-genomes and pan-genomes of each group. To compare the core-genomes and pan-genomes between different groups, a comparative analysis was performed focusing on specific KEGG pathways and genes associated with environmental signal response, critical substance synthesis, and metabolism (Figure 4). The bar charts in different colors represent the core-genome and pan-genome differences between Group 2 and Group 3 + Group 4, as well as Group 3 + Group 4 and Group 5 (Group 1, with a small number of strains, is not discussed in this section). To assess the impact of changes in gene numbers on each KEGG pathway, the relative enrichment ratio (RER) was calculated using RER values of 0.1 and 0.2 as screening criteria. The effective enrichment pathways from the comparative core-genomic and pan-genomic analysis between different groups are presented in Figure 4 (Supplementary Table S4). Although the pan-genome contains the core-genome, differences were observed between the core-genomes and pan-genomes of the two adjacent groups. This is because our analysis method removes the influence of the number of metabolic pathway background genes. In the pan-genome of Group 2 and not in Group 3 + Group 4, enriched pathways included others, chlorocyclohexane and chlorobenzene degradation, xylene degradation, and some metabolism-related pathways (especially disinfectant substances). On the other hand, pathways such as bacterial chemotaxis, flagellar assembly, bacterial secretion system, biofilm formation, replication and repair, and nutrient metabolism-related pathways were mainly enriched in the pan-genome of Group 3 + Group 4 and not in Group 2. In the core-genome, pathways such as bacterial chemotaxis and biofilm formation were enriched in Group 3 + Group 4 and not in Group 2, while pathways related to transport, drug metabolism, and some metabolism-related pathways were found in the core-genome of Group 2 and not in Group 3 + Group 4. In the pan-genome of Group 3 + Group 4 and not in Group 5, enriched pathways included bacterial chemotaxis, flagellar assembly, bacterial secretion system,



Core genes in Group2 and not in Group3+Group4

	Num	RER
Streptomycin biosynthesis	4	0.19
Drug metabolism - cytochrome P450	4	0.16
Drug metabolism - other enzymes	5	0.16
Bacterial motility proteins	20	0.13
Platinum drug resistance	6	0.13
Transport	6	0.13
Metabolism of xenobiotics by cytochrome P450	4	0.12
Tyrosine metabolism	9	0.11
Monobactam biosynthesis	3	0.11
Valine, leucine and isoleucine biosynthesis	2	0.11
Chemical carcinogenesis - DNA adducts	3	0.10

Pan genes in Group2 and not in Group3+Group4

	Num	RER
Others	3	0.30
Chlorocyclohexane and chlorobenzene degradation	10	0.24
Xylene degradation	7	0.21
Dioxin degradation	5	0.16
Toluene degradation	6	0.13
Biotin metabolism	3	0.13
Geraniol degradation	2	0.13
Benzoate degradation	13	0.12
Fatty acid biosynthesis	5	0.12
Styrene degradation	3	0.12
Drug metabolism - cytochrome P450	3	0.12

Core genes in Group3+Group4 and not in Group2

	Num	RER
Bacterial chemotaxis	7	0.27
Global maps only	1	0.25
Biofilm formation - Pseudomonas aeruginosa	12	0.13

Pan genes in Group3+Group4 and not in Group2

	Num	RER
Bacterial chemotaxis	14	0.54
Flagellar assembly	20	0.36
Limonene and pinene degradation	4	0.33
Bacterial secretion system	23	0.31
Bacterial motility proteins	42	0.28
Biofilm formation - Pseudomonas aeruginosa	23	0.26
Caprolactam degradation	5	0.23
Secretion system	66	0.15
C5-Branched dibasic acid metabolism	4	0.14
Drug metabolism - cytochrome P450	3	0.12
Replication and repair	7	0.12
Carbohydrate metabolism	3	0.12
Fluorobenzoate degradation	2	0.11
Valine, leucine and isoleucine biosynthesis	2	0.11
O-Antigen nucleotide sugar biosynthesis	10	0.10
Quorum sensing	29	0.10
Fatty acid degradation	6	0.10
Valine, leucine and isoleucine degradation	7	0.10
Others	1	0.10
Glyoxylate and dicarboxylate metabolism	10	0.10

Pan genes in Group3+Group4 and not in Group2

	Num	RER
Bacterial chemotaxis	10	0.38
Flagellar assembly	20	0.36
Limonene and pinene degradation	4	0.33
Bacterial motility proteins	37	0.24
Replication and repair	13	0.22
Valine, leucine and isoleucine biosynthesis	4	0.21
Bacterial secretion system	14	0.19
Caprolactam degradation	4	0.18
C5-Branched dibasic acid metabolism	5	0.17
Biofilm formation - Pseudomonas aeruginosa	15	0.17
Drug metabolism - cytochrome P450	4	0.16
Biofilm formation - Escherichia coli	8	0.13
Alanine, aspartate and glutamate metabolism	9	0.13
Vitamin B6 metabolism	3	0.13
Secretion system	51	0.12
Carbohydrate metabolism	3	0.12
Mismatch repair	5	0.11
Insect hormone biosynthesis	2	0.11
Fluorobenzoate degradation	2	0.11
Nucleotide excision repair	5	0.11
Others	1	0.10
Cell cycle - Caulobacter	3	0.10

Core genes in Group3+Group4 and not in Group2

	Num	RER
Biofilm formation - Pseudomonas aeruginosa	9	0.10

Core genes in Group5 and not in Group3+Group4

	Num	RER
Furfural degradation	4	0.50
Global maps only	1	0.25
Valine, leucine and isoleucine biosynthesis	3	0.16
Tyrosine metabolism	11	0.13
Transport	6	0.13
Bacterial motility proteins	19	0.13
Platinum drug resistance	5	0.11
C5-Branched dibasic acid metabolism	3	0.10
Others	1	0.10
Streptomycin biosynthesis	2	0.10

Pan genes in Group5 and not in Group3+Group4

	Num	RER
Toluene degradation	11	0.24
Xylene degradation	7	0.21
Others	2	0.20
Chlorocyclohexane and chlorobenzene degradation	8	0.20
Dioxin degradation	4	0.13
Styrene degradation	3	0.12
Benzoate degradation	12	0.11
Phenylalanine metabolism	8	0.11
Bacterial secretion system	7	0.10

FIGURE 4
Screen results of comparative core- and pan-genomic analysis between different Groups. The chart displays a bar chart showing the number of genes and a line chart representing the Relative Enrichment Ratio (RER) for each KEGG pathway. The RER is a measure of the correlation between gene enrichment and the impact on the corresponding pathway. On the right side of the chart, the screen result of the comparative core- and pan-genomic analysis is presented. To provide visual clarity, KEGG pathways with RER values between 0.1 and 0.20 are labeled in black, while pathways with RER values greater than 0.20 are labeled in dark red.

biofilm formation, caprolactam degradation, limonene and pinene degradation, valine, leucine, and isoleucine biosynthesis, and nutrient metabolism-related pathways. Pathways such as furfural degradation, transport, and some metabolism-related pathways were enriched in the core-genome of Group 5 and not in Group 3 + Group 4. The pan-genome of Group 5 showed enrichment in pathways such as chlorocyclohexane and chlorobenzene degradation, others, and some metabolism-related pathways (especially disinfectant substances) that were not found in Group 3 + Group 4. These differences in core-genomes and pan-genomes between different groups reveal the genomic characteristics formed through the adaptive evolution of *Ralstonia pickettii* under different environmental conditions. Additionally, these results indicate the diversity of the *Ralstonia pickettii* genome.

The key mutation accumulation of *Ralstonia pickettii* during the adaptive evolution from complex environmental conditions to drinking water

Specific gene mutations play a crucial role in promoting adaptive evolution in bacteria and can lead to genetic changes that help microorganisms adapt to diverse environmental conditions (Levin and Bergstrom, 2000; Chattopadhyay et al., 2013). To analyze the gene mutations associated with adaptive evolution in *Ralstonia pickettii* in drinking water, we examined the distribution of single nucleotide polymorphisms (SNPs) loci and their functional annotations in key branch-points on the phylogenetic tree (Figure 5A; Supplementary Table S5). As shown in Figure 5A, it can be observed that SNP loci occurred at branch-point I was associated with 777 orthologous gene families. Functional enrichment analysis based on KEGG annotation revealed several pathways that were affected by SNP loci. Orthologous gene families with a high number of SNP loci ($n \geq 20$) were associated with pathways, such as glycerolipid metabolism, glycerophospholipid metabolism, valine, leucine, and isoleucine biosynthesis, pantothenate and CoA biosynthesis, two-component system, pertussis, protein kinases, and lipid biosynthesis proteins. Notably, there were 84 SNP loci found in orthologous gene families associated with purine metabolism. Orthologous gene families with a moderate number of SNP loci ($10 \leq n < 20$) were linked to pathways, such as transporter, two-component system, peptidases and inhibitors, protein export, glycine, serine, and threonine metabolism, carbon fixation in photosynthetic organisms, pyruvate metabolism, galactose metabolism, and glycolysis/gluconeogenesis. Similarly, orthologous gene families with a lower number of SNP loci ($5 \leq n < 10$) were also associated with pathways, such as transporter and two-component system. Even genes with fewer accumulated SNPs ($1 < n < 5$) were primarily distributed in pathways related to transporter, two-component system, quorum sensing, enzymes with EC numbers, and bacterial motility proteins. At branch-point II, SNP loci were associated with 37 orthologous gene families. The functional enrichment analysis based on KEGG annotation indicated that orthologous gene families with a lower number of SNP loci ($1 < n < 5$) were associated with the pathway transporters. The presence of SNP loci at specific branch-points in the phylogenetic tree suggests their role in shaping specific branches and clusters in population genetics analysis. Thus, the distribution of these SNP loci can provide insights into the accumulation of mutations during the adaptive evolution of *Ralstonia pickettii* under complex genetic backgrounds. These mutations are likely to affect specific physiological functions, providing the organism with the potential ability to survive in drinking water.

Genetic characteristics of *Ralstonia pickettii* in the process of adaptive evolution in drinking water

Drinking water creates a unique environmental pressure on *Ralstonia pickettii*, leading to potential variations in its genetic characteristics during the process of adaptive evolution. To analyze these variations, we conducted a genomic analysis of strains isolated from drinking water in Group 2 and Group 5 by COG annotation (Figure 5C; Supplementary Table S6). As shown in Figure 5C, strains isolated from drinking water in the Group 2 exhibited a higher number of core orthologous gene families compared with the Group 5. However, strains from drinking water in the Group 5 displayed a larger number of accessory and unique orthologous gene families. These findings may be attributed to HGT events occurring between *Ralstonia pickettii* and other bacteria within the same environmental niche. It is worth noting that strains of *Burkholderia* have also been isolated from the potable water system in the International Space Station (O'Rourke et al., 2020), indicating the occurrence of HGT events in similar environments. Additionally, the strains isolated from drinking water in the Group 5 had a higher number of HGT genes compared with other strains in this study, further supporting the notion of genetic exchange through HGT. This aligns with the previous results, where strains from the Group 5 exhibited more HGT genes than other strains in this research. To analyze the gene mutations accumulated during the adaptive evolution of *Ralstonia pickettii* in drinking water, we examined the distribution of SNP loci in strains isolated from drinking water (Figure 5B). In the Group 2, SNP loci were identified in four orthologous gene families. Two orthologous gene families had a small number of SNP loci ($1 < n < 5$) and were associated with KEGG pathways, such as pinene, camphor, and geraniol degradation, DNA repair and recombination proteins, and microbial metabolism in diverse environments. One orthologous gene family had a moderate number of SNP loci ($5 \leq n < 10$) and was involved in various KEGG pathways, including the citrate cycle (TCA cycle), oxidative phosphorylation, butanoate metabolism, carbon fixation pathways in prokaryotes, biosynthesis of secondary metabolites, microbial metabolism in diverse environments, carbon metabolism, and legionellosis. One orthologous gene family had a high number of SNP loci ($n \geq 20$) and was linked to the KEGG pathway, such as linoleic acid metabolism. In the Group 5, SNP loci were distributed across 58 orthologous gene families. Fifty orthologous gene families had a small number of SNP loci ($1 < n < 5$) and were mainly associated with carbohydrate metabolism, energy metabolism, transporters, and quorum sensing. Five orthologous gene families had a moderate number of SNP loci ($5 \leq n < 10$) and were involved in KEGG pathways, such as replication and repair, legionellosis, transporters, and an unknown pathway (TnsA endonuclease, transcriptional regulator). Three orthologous gene families had a high number of SNP loci ($n \geq 20$) and were associated with KEGG

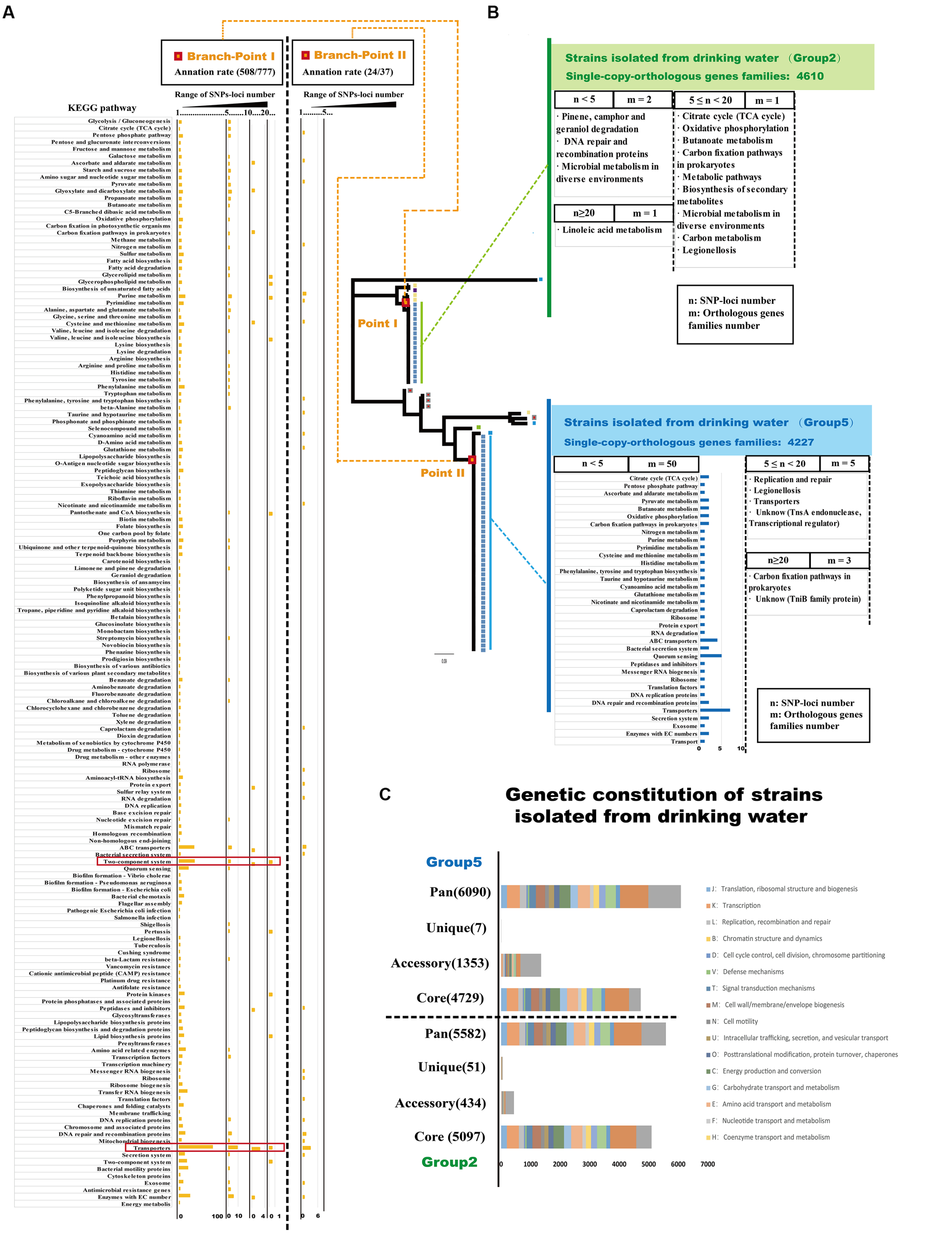


FIGURE 5
SNPs-loci distribution reveals the adaptive evolutionary characteristic of *Ralstonia pickettii* in drinking water. **(A)** The distribution of SNPs-loci at the two key branch-points of the phylogenetic tree is depicted. The bar chart represents the enrichment of orthologous gene families with SNPs-loci after KEGG annotation. The results are divided into three ranges based on the number of SNPs-loci (1 < *n* < 5, 5 ≤ *n* < 10, 10 ≤ *n* < 20, *n* ≥ 20), and the corresponding KEGG pathways are labeled on the left. **(B)** The distribution of SNPs-loci in *Ralstonia pickettii* strains isolated from drinking water is shown. The results are divided into three ranges based on the number of SNPs-loci (1 < *n* < 5, 5 ≤ *n* < 20, *n* ≥ 20), and different groups are color-coded (green and blue). The bar chart represents the function enrichment of orthologous gene families with SNPs-loci after KEGG annotation. **(C)** The genomic characteristics and COG annotation of strains isolated from drinking water in Group 2 and Group 5 are presented.

pathways, including carbon fixation pathways in prokaryotes and an unknown pathway (TniB family protein: recombinase). Overall, the analysis suggests that under the environmental pressure of drinking water, *Ralstonia pickettii* accumulates mutations in orthologous gene families associated with linoleic acid metabolism, carbon fixation pathways in prokaryotes, and DNA recombination and repair. Notably, many orthologous gene families that accumulate mutations are related to carbon fixation pathways associated with the energy metabolism of *Ralstonia pickettii*.

Potential pathogenicity and drug resistance of *Ralstonia pickettii*

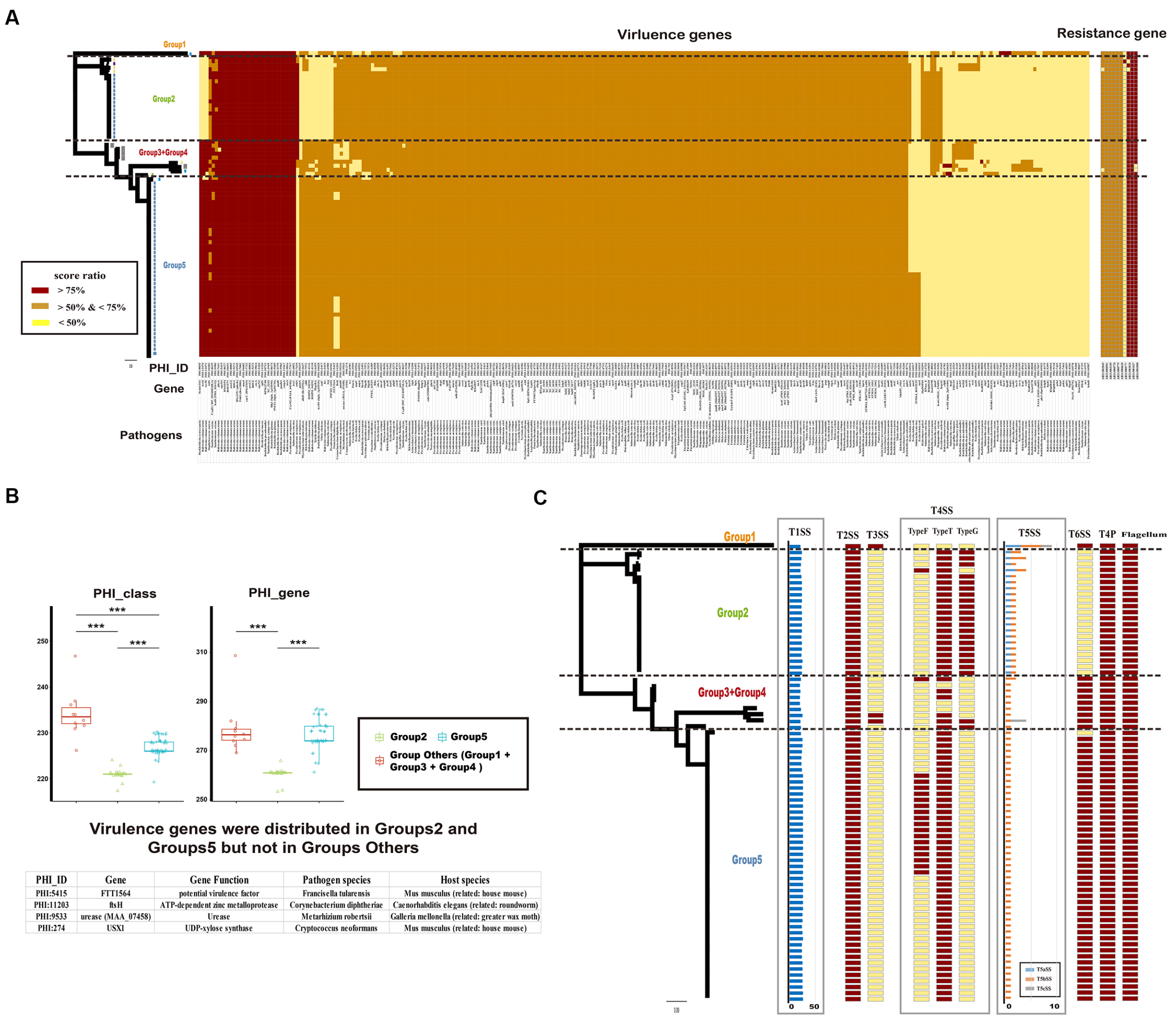
To identify potential pathogenic characteristics, the virulence genes of *Ralstonia pickettii* were investigated. A total of 285 gene families were identified that matched with virulence genes in the PHI database (Figure 6A and the detail information is shown in Supplementary Table S7). Among these gene families, 213 (74.7%) were shared by almost all *Ralstonia pickettii* strains, indicating a potential shared pathogenic capacity. Additionally, it was found that all *Ralstonia pickettii* strains possess the general secretory pathway (GSP) from *Ralstonia solanacearum*. The study also examined the role of virulence genes in the adaptive evolution of *Ralstonia pickettii* in drinking water. It was observed that strains in Group 1 + Group 3 + Group 4 had a significantly higher number of virulence genes (234.2 ± 10.5) compared with other groups. In contrast, strains of the Group 2 had fewer virulence genes (262.8 ± 24.0) and PHI classes (222.1 ± 13.0) compared with the other groups (Figure 6B). Specific virulence genes mainly distributed in the Groups 2 and 5 were identified, including genes associated with potential virulence factors from other bacteria. These genes may be related to the environmental adaptability of *Ralstonia pickettii* in drinking water.

Furthermore, the presence of macromolecular secretion systems, such as type I (T1SS), type II (T2SS), type III (T3SS), type IV (T4SS), type V (T5SS), and type VI (T6SS) secretion systems, as well as type IV pilus (T4P), flagellum, and Tad pilus, was investigated (Figure 6C; Supplementary Table S8). It was found that all *Ralstonia pickettii* strains possessed T1SS, T2SS, T4P, and flagellum. T3SS was present only in certain strains of Group 1 and Group 3 + 4, indicating a relatively strong virulence in these groups. Different types of T4SS and T5SS were also identified in *Ralstonia pickettii* strains, with variations in their distribution among different groups. Additionally, we examined the genotypic profiles of antimicrobial resistance in *Ralstonia pickettii*. It was found that almost all strains possessed eight antimicrobial resistance genes: ARO:3003049 (*rosB*), ARO:3003010 (*ceoB*), ARO:3000778 (*adeG*), ARO:3000779 (*adeH*), ARO:3001417 (*OXA-22*), ARO:3001808 (*OXA-60*), ARO:3000501 (*rpoB2*), and ARO:3003105 (*dfrA3*), which was associated with resistance-nodulation-cell division (RND) antibiotic efflux pump, major facilitator superfamily (MFS) antibiotic efflux pump, beta-lactamase, and trimethoprim-resistant dihydrofolate reductase. These resistance genes may confer corresponding clinical resistance phenotypes to *Ralstonia pickettii*. Overall, the analysis of virulence genes, macromolecular secretion systems, and antimicrobial resistance genes provides insights into the potential pathogenic characteristics and adaptive evolution of *Ralstonia pickettii* in drinking water environments.

Discussion

Ralstonia pickettii, being a versatile pathogen, can be found in various habitats such as water conditions, sediments, plants, and drinking water. As a generalist, *Ralstonia pickettii* likely maintains large genomes with a higher number of functional genes to adapt to diverse environmental conditions (Thomas et al., 2016). The presence of an open pan-genome with variable gene families plays a crucial role in the genome plasticity and genetic diversity of *Ralstonia pickettii*, which is significant for its adaptive evolution (Pandit et al., 2009; Sriswasdi et al., 2017). The analysis conducted in this study revealed the potential role of both the core-genome and accessory genes in the adaptive evolution of *Ralstonia pickettii*. The core genome, which consists of genes shared by all strains, was found to be enriched in categories related to substance synthesis, nutrient transport, and metabolism (COG-J, −E, -F, -H, and -I). These essential genes enable *Ralstonia pickettii* to efficiently acquire nutrients from the environment and adapt metabolically to occupy different ecological niches (Goyal, 2018; Iranzo et al., 2019; Cummins et al., 2022). On the other hand, the accessory genome, which consists of genes present in some but not all strains, was enriched in categories associated with bacterial proliferation, motility, and substance transport (COG-L, -N, and -U). These genes contribute to maintaining normal physiological functions in bacteria and provide an evolutionary dynamic to keep the pan-genome open under different environmental stresses (Garcia-Garcera and Rocha, 2020; Maistrenko et al., 2020).

Phylogenetic and genetic population analyses have divided *Ralstonia pickettii* into five distinct groups. Strains isolated from drinking water in the International Space Station (ISS) were primarily distributed in the Group 2 and Group 5, while strains associated with humans (*Homo sapiens*) were found in the Group 3 and Group 4. Comparative analysis of the core- and pan-genomes among these groups has revealed the diverse genomic characteristics of *Ralstonia pickettii* during its adaptive evolution under different environmental conditions. The pan-genome of strains in the Group 2 or Group 5 exhibited a high abundance of genes associated with the metabolism of specific substances, particularly those involved in the degradation of common disinfectants. Pathways such as chlorocyclohexane and chlorobenzene degradation, xylene degradation, dioxin degradation, toluene degradation, and benzoate degradation were prominently represented. Additionally, the core-genome of strains in the Group 2 or Group 5 contained genes associated with pathways related to transport, drug metabolism, and certain amino acid metabolism. However, these genes were not found in the Group 3 + Group 4 strains. Importantly, it is worth noting that most strains in the Group 2 and Group 5 were isolated from the potable water system in the International Space Station, where current physical and chemical methods are generally effective at maintaining water cleanliness. Bacteria residing in the potable water system are constantly exposed to extreme environmental pressures from drugs, disinfectants, and oligotrophic conditions (Wong et al., 2010; Bornemann et al., 2015; Thornhill and Kumar, 2018). The unique core- and pan-genomes found in strains from the Group 2 and Group 5 likely contribute to the environmental adaptability of *Ralstonia pickettii*, enabling its survival in the potable water system of the International Space Station. On the



Virulence genes were distributed in Groups2 and Groups5 but not in Groups Others

PHI_ID	Gene	Gene Function	Pathogen species	Host species
PHI:5415	FTT1564	potential virulence factor	Francisella tularensis	Mus musculus (related: house mouse)
PHI:11203	ftsH	ATP-dependent zinc metalloprotease	Corynebacterium diphtheriae	Caenorhabditis elegans (related: roundworm)
PHI:9533	urease (MAA_07458)	Urease	Metarhizium robertsii	Galleria mellonella (related: greater wax moth)
PHI:274	USX1	UDP-xylose synthase	Cryptococcus neoformans	Mus musculus (related: house mouse)

FIGURE 6
Potential pathogenicity and drug resistance of *Ralstonia pickettii* strains. **(A)** The heatmap displays the distribution of virulence genes based on the PHI database and resistance genes based on the CARD database. Each gene is color-coded based on the score ratio from the blast result of each genome. The PHI-ID, gene name, and associated pathogen are labeled below the heatmap. **(B)** The box chart shows the difference in virulence genes among different groups (Group 2, Group 5, and Group Others: Group 1 + Group 3 + Group 4). The chart indicates the number of virulence genes and PHI classes for each group. Statistical significance is indicated by asterisks (****t*-test *p*-value < 0.001). Virulence genes that are exclusively present in Group 2 and Group 5 are listed at the bottom. **(C)** The distribution of macromolecular secretion systems in *Ralstonia pickettii* is represented. Dark red boxes indicate the presence of a macromolecular system within a genome, while yellow boxes indicate the absence of a macromolecular system.

other hand, the pan-genome and core-genome specific to strains in the Group 3 + Group 4 were primarily associated with pathways such as bacterial chemotaxis, flagellar assembly, bacterial secretion system, biofilm formation, and nutrient metabolism-related pathways. These genes potentially provide *Ralstonia pickettii* with enhanced motility, stress resistance, and the ability to respond to and utilize diverse nutrients, thereby achieving better environmental adaptability under complex environmental conditions (Tolker-Nielsen, 2015; Goyal, 2018; Rossi et al., 2018; Colin et al., 2021). Furthermore, the differences in the pan-genomic composition among different groups highlight the close relationship between prokaryotic non-core genes and various evolutionary models of population structure and dynamics. Gene composition evolution through gene gain and loss tends to occur at a faster rate than sequence evolution (Iranzo et al., 2019).

Genetic drift and gene conversion are crucial processes in the adaptive evolution of bacteria, contributing to genetic plasticity and genomic changes (Power et al., 2021). Our analysis revealed significant differences in the number of genomic islands (GIs), prophages, and horizontally transferred genes (HGT) among the different groups (Figure 3B), indicating the role of genetic plasticity and genomic evolution driven by mobile genetic elements and HGT in *Ralstonia pickettii*. The functional differences observed in GI genes and HGT genes among strains in different groups suggest that these processes introduce new characteristics during bacterial adaptation (Rosenberg, 2001). Specifically, we observed that GI genes in the Group 2 and Group 5 were significantly enriched in COG-B and COG-M functional categories, while HGT genes in these groups were significantly enriched in COG-L. Similar functional enrichments were observed when comparing GI genes and HGT genes between strains isolated from drinking water and other environments. The low osmotic pressure in drinking water imposes greater environmental stress on bacteria, necessitating the presence of genes involved in envelope stress response. Additionally, gene families associated with DNA replication, recombination, and repair (COG-L) likely play crucial roles in the adaptive plasticity of the bacterial genome by allowing individuals to adjust their recombination and mutation rates (Poolman et al., 2002; Pandit et al., 2009; Paul, 2013; Milner et al., 2023). Our study demonstrates that *Ralstonia pickettii* may acquire important genes through GIs and HGT to achieve environmental adaptation in water environments.

During adaptive evolution, bacteria undergo mutations in key orthologous gene families, leading to the emergence of new branches in phylogenetic trees (Wellenreuther et al., 2019; Wei et al., 2022). In our study, we focused on two branch-points associated with strains isolated from drinking water. We observed that orthologous gene families associated with pathways such as transporters and two-component systems were more prone to accumulating mutations and forming single nucleotide polymorphisms (SNPs). Transporter proteins are considered ecological assets and features of microbial pangenomes, capable of providing niche-defining phenotypes (Wellenreuther et al., 2019; Wei et al., 2022; Milner et al., 2023). Furthermore, during adaptive evolution in drinking water, *Ralstonia pickettii* accumulated specific gene mutations in key orthologous gene families related to linoleic acid metabolism, carbon fixation pathways in prokaryotes, and DNA recombination and repair. Notably, many of the orthologous gene families showing mutations were associated with carbon fixation pathways, indicating that the carbon metabolism and energy metabolism of *Ralstonia pickettii* are significantly influenced by environmental pressures during adaptive evolution in drinking water. These mutations are likely to impact related physiological functions and provide insights into the direction of adaptive evolution. On the other hand, microgravity also serves as an environmental factor for bacteria to live in the ISS. It has been reported that microgravity can enhance certain physiological functions of bacteria, such as growth, biofilm formation, virulence, and antibiotic resistance. However, most of these physiological changes are related to the regulation of gene expression (Singh et al., 2018; Vaishampayan et al., 2018; Vaishampayan and Grohmann, 2019). The potential impact of microgravity on this process can be discussed in future studies.

The virulence genotypic profiles and the distribution of macromolecular secretion systems revealed potential pathogenic characteristics of bacteria (Abby et al., 2014; Abby and Rocha, 2017; Urban et al., 2020). Most virulence genes (74.7%) are shared by all *Ralstonia pickettii* strains, indicating a potential shared pathogenicity. We focused on the virulence genes only primarily distributed in the Group 2 and Group 5, which may be related to the environmental adaptability of bacteria in drinking water. Based on the sequence alignment results, these genes associated with FTT1564 (potential virulence factor from *Francisella tularensis*), FtsH (ATP-dependent zinc metalloprotease *Corynebacterium diphtheria*), MAA_07458 (urease from *Metarhizium robertsii*), and USX1 (UDP-xylose synthase from *Cryptococcus neoformans*). The role of these virulence genes in the environmental adaptability of *Ralstonia pickettii* in drinking water needs further study. Macromolecular secretion systems, including T4SS, three types of T5SS (T5aSS, T5bSS, and T5cSS), and T6SS, reflected differences in potential pathogenicity among the different groups which were found in *Ralstonia pickettii*. The unique distribution of T5SS may significantly impact the potential virulence of different groups (Fan et al., 2016), but it may not contribute to environmental adaptation. T6SS has been reported to contribute to bacterial pathogenesis by translocating substrates in the host and facilitating competition with other bacteria in their niches (Grohmann et al., 2018). In *Ralstonia pickettii*, T6SS was found in all groups except the Group 2. The specific distribution of different macromolecular secretion systems among *Ralstonia pickettii* genomes suggests their potential pathogenicity under diverse environmental conditions. This suggests that these macromolecular secretion systems play a role in the pathogenicity of the *Ralstonia pickettii*. In addition, the antimicrobial resistance genotypic profiles of *Ralstonia pickettii* revealed the presence of eight antimicrobial resistance genes and were shared with almost all strains. OXA-60 and OXA-22 were two chromosomal resistance genes in *Ralstonia pickettii* (Nordmann et al., 2000; Girlich et al., 2004). Resistance-nodulation-cell division (RND) family has been reported as the dominant intrinsic and acquired multidrug resistance mechanism in *Burkholderia* species (Podnecky et al., 2015). The results indicated that these resistance genes, as the core gene, did not change with the evolution of the *Ralstonia pickettii* genome, and contributed to the clinical resistance phenotype, which highlighted the importance of addressing antimicrobial resistance in *Ralstonia pickettii* infections.

In conclusion, the comprehensive genomic analysis has provided valuable insights into the genetic diversity, potential pathogenicity, and adaptive evolution of *Ralstonia pickettii*. The study has highlighted the significance of virulence-related elements and antimicrobial resistance genes in the pathogenicity of bacteria. Moreover, the analysis has provided a deeper understanding of the genetic characteristics and adaptive evolutionary processes of *Ralstonia pickettii* in the context of the drinking water environment. These findings contribute to our knowledge of this bacterium and have implications for addressing its pathogenic potential and antimicrobial resistance in healthcare and environmental settings. Further research in this area will continue to enhance our understanding of *Ralstonia pickettii* and its interactions with its environment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

CY: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. TA: Data curation, Resources, Software, Writing – original draft. XL: Data curation, Resources, Software, Writing – review & editing. JZ: Data curation, Resources, Software, Writing – review & editing. ZL: Data curation, Resources, Software, Writing – review & editing. JG: Data curation, Resources, Software, Writing – review & editing. RH: Data curation, Investigation, Writing - review & editing. ZF: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the National Key Research and Development Program of China (2021YFA1301202), National Natural Science Foundation of China (NSFC) Program (grant no. 82273676), and the Tianjin Municipal Education Commission (no. 2020KJ202).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1272636/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Comparative core- and pan-genomic analysis between different Groups based on KEGG annotation. The histogram represented the gene-number of KEGG categories and the color of each Group corresponded to the color in Figure 1.
Detailed information on the KEGG categories was shown on the bottom.

SUPPLEMENTARY TABLE S1
Detailed strain-information and genetic characteristics of the strains in this study.

SUPPLEMENTARY TABLE S2
Detailed information of pan-blast-screen results.

SUPPLEMENTARY TABLE S3
Detailed information of MGEs and HGT genes.

SUPPLEMENTARY TABLE S4
Screen and statistics results of comparative core-and pan-genome of each Group after KEGG annotation analysis.

SUPPLEMENTARY TABLE S5
Detailed information of the distribution of SNPs-loci at key branch-point.

SUPPLEMENTARY TABLE S6
Detailed information of mutation accumulation during adaption evolution in drinking water.

SUPPLEMENTARY TABLE S7
Detailed information of PHI-screen-results and corresponding PHI genes.

SUPPLEMENTARY TABLE S8
Detailed information of CARD-screen-results and the distribution of macromolecular secretion systems.

SUPPLEMENTARY TABLE S9
Screen and statistics results of core-and pan-genome of each Group after KEGG annotation analysis.

References

Abby, S. S., Néron, B., Ménager, H., Touchon, M., and Rocha, E. P. C. (2014). MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PloS One* 9, e110726–e110729. doi: 10.1371/journal.pone.0110726

Abby, S. S., and Rocha, E. P. C. (2017). Identification of protein secretion systems in bacterial genomes using MacSyFinder. *Methods Mol. Biol.* 1615, 1–21. doi: 10.1007/978-1-4939-7033-9_1

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387

Baker, M. A., Rhee, C., Tucker, R., Vaidya, V., Holtzman, M., Seethala, R. R., et al. (2022). *Ralstonia pickettii* and *Pseudomonas aeruginosa* bloodstream infections associated with contaminated extracorporeal membrane oxygenation water heater devices. *Clin. Infect. Dis.* 75, 1838–1840. doi: 10.1093/cid/ciac379

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing GLau, B. Y., Hoad, G., et al. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45, W30–W35. doi: 10.1093/nar/gkx343

Bornemann, G., Waßer, K., Tonat, T., Moeller, R., Bohmeier, M., and Hauslage, J. (2015). Natural microbial populations in a water-based biowaste management system for space life support. *Life Sci. Space Res.* 7, 39–52. doi: 10.1016/j.lssr.2015.09.002

Bottacini, F., Medini, D., Pavesi, A., Turroni, F., Foroni, E., Riley, D., et al. (2010). Comparative genomics of the genus *Bifidobacterium*. *Microbiology* 156, 3243–3254. doi: 10.1099/mic.0.039545-0

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, Orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Chattopadhyay, S., Paul, S., Dykhuizen, D. E., and Sokurenko, E. V. (2013). Tracking recent adaptive evolution in microbial species using TimeZone. *Nat. Protoc.* 8, 652–665. doi: 10.1038/nprot.2013.031

Chen, C. M., Liu, J. J., Chou, C. W., Lai, C. H., and Wu, L. T. (2015). RpA, an extracellular protease similar to the metalloprotease of serralysin family, is required for pathogenicity of *Ralstonia pickettii*. *J. Appl. Microbiol.* 119, 1101–1111. doi: 10.1111/jam.12903

Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M., and Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* 30, 1224–1228. doi: 10.1093/molbev/mst028

Colin, R., Ni, B., Laganenka, L., and Sourjik, V. (2021). Multiple functions of flagellar motility and chemotaxis in bacterial physiology. *FEMS Microbiol. Rev.* 45:fuab038. doi: 10.1093/femsre/fuab038

Cummins, E. A., Hall, R. J., McInerney, J. O., and McNally, A. (2022). Prokaryote pangenomes are dynamic entities. *Curr. Opin. Microbiol.* 66, 73–78. doi: 10.1016/j.mib.2022.01.005

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Euzéby, J. (1991). List of prokaryotic names with standing in nomenclature. Available at: https://www.bacterio.net/ralstonia.html Accessed January 2021.

Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., et al. (2014). Mobile elements drive recombination hotspots in core-genome of *Staphylococcus aureus*. *Nat. Commun.* 5:3956. doi: 10.1038/ncomms4956

Fan, E., Chauhan, N., Udatha, D. B. R. K. G., Leo, J. C., and Linke, D. (2016). Type V secretion Systems in Bacteria. *Microbiol Spectr.* 4:2015. doi: 10.1128/microbiolspec.VMBF-0009-2015

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* 43, D261–D269. doi: 10.1093/nar/gku1223

Garcia-Garcera, M., and Rocha, E. P. C. (2020). Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat. Commun.* 11:758. doi: 10.1038/s41467-020-14572-x

Girlich, D., Naas, T., and Nordmann, P. (2004). OXA-60, a chromosomal, inducible, and imipenem-hydrolyzing class D beta-lactamase from *Ralstonia pickettii*. *Antimicrob. Agents Chemother.* 48, 4217–4225. doi: 10.1128/AAC.48.11.4217-4225.2004

Goyal, A. (2018). Metabolic adaptations underlying genome flexibility in prokaryotes. *PLoS Genet.* 14:e1007763. doi: 10.1371/journal.pgen.1007763

Grohmann, E., Christie, P. J., Waksman, G., and Backert, S. (2018). Type IV secretion in gram-negative and gram-positive bacteria. *Mol. Microbiol.* 107, 455–471. doi: 10.1111/mmi.13896

Gyles, C., and Boerlin, P. (2014). Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet. Pathol.* 51, 328–340. doi: 10.1177/0300985813511131

Heaps, H. S. (1978). *Information retrieval – Computational and theoretical aspects*. Academic Press, Orlando, FL

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Iranzo, J., Wolf, Y. I., Koonin, E. V., and Sela, I. (2019). Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat. Commun.* 10:5376. doi: 10.1038/s41467-019-13429-2

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Levin, B. R., and Bergstrom, C. T. (2000). Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6981–6985. doi: 10.1073/pnas.97.13.6981

Li, J., Yao, Y., Xu, H. H., Hao, L., Deng, Z., Rajakumar, K., et al. (2015). SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environ. Microbiol.* 17, 2196–2202. doi: 10.1111/1462-2920.12794

Liu, L., Tai, C., Bi, D., Ou, H.-Y., Rajakumar, K., and Deng, Z. (2012). SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.* 41, D660–D665. doi: 10.1093/nar/gks1248

Liu, X., Zarfel, G., van der Weijden, R., Loiskandl, W., Bitschnau, B., Dinkla, I. J. T., et al. (2021). Density-dependent microbial calcium carbonate precipitation by drinking water bacteria via amino acid metabolism and biosorption. *Water Res.* 202:117444. doi: 10.1016/j.watres.2021.117444

Maistrenko, O. M., Mende, D. R., Luetge, M., Hildebrand, F., Schmidt, T. S. B., Li, S. S., et al. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* 14, 1247–1259. doi: 10.1038/s41396-020-0600-z

Menekşe, Ş., Hacıseyitoğlu, D., Süzük Yıldız, S., and Bayrakdar, F. (2022). An outbreak of *Ralstonia pickettii* bloodstream infection and clinical outcomes. *J. Infect. Dev. Ctries.* 16, 705–711. doi: 10.3855/jidc.15159

Milner, D. S., Galindo, L. J., Irwin, N. A. T., and Richards, T. A. (2023). Transporter proteins as ecological assets and features of microbial eukaryotic Pangenomes. *Annu. Rev. Microbiol.* 77, 45–66. doi: 10.1146/annurev-micro-032421-115538

Nasir, N., Sayeed, M. A., and Jamil, B. (2019). *Ralstonia pickettii* bacteremia: an emerging infection in a tertiary care hospital setting. *Cureus.* 11:e5084. doi: 10.7759/cureus.5084

Nordmann, P., Poirel, L., Kubina, M., Casetta, A., and Naas, T. (2000). Biochemical-genetic characterization and distribution of OXA-22, a chromosomal and inducible class D beta-lactamase from *Ralstonia* (*Pseudomonas*) *pickettii*. *Antimicrob. Agents Chemother.* 44, 2201–2204. doi: 10.1128/AAC.44.8.2201-2204.2000

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304. doi: 10.1038/35012500

O'Rourke, A., Lee, M. D., Nierman, W. C., Everroad, R. C., and Dupont, C. L. (2020). Genomic and phenotypic characterization of *Burkholderia* isolates from the potable water system of the international Space Station. *PloS One* 15:e0227152. doi: 10.1371/journal.pone.0227152

Pandit, S. N., Kolasa, J., and Cottenie, K. (2009). Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* 90, 2253–2262. doi: 10.1890/08-0851.1

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

Paul, D. (2013). Osmotic stress adaptations in rhizobacteria. *J. Basic Microbiol.* 53, 101–110. doi: 10.1002/jobm.201100288

Podnecky, N. L., Rhodes, K. A., and Schweizer, H. P. (2015). Efflux pump-mediated drug resistance in *Burkholderia*. *Front. Microbiol.* 6:305. doi: 10.3389/fmicb.2015.00305

Poolman, B., Blount, P., Folgering, J. H., Friesen, R. H., Moe, P. C., and van der Heide, T. (2002). How do membrane proteins sense water stress? *Mol. Microbiol.* 44, 889–902. doi: 10.1046/j.1365-2958.2002.02894.x

Power, J. J., Pinheiro, F., Pompei, S., Kovacova, V., Yüksel, M., Rathmann, I., et al. (2021). Adaptive evolution of hybrid bacteria by horizontal gene transfer. *Proc. Natl. Acad. Sci. U. S. A.* 118:118. doi: 10.1073/pnas.2007873118

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

Riley, P. S., and Weaver, R. E. (1975). Recognition of *Pseudomonas pickettii* in the clinical laboratory: biochemical characterization of 62 strains. *J. Clin. Microbiol.* 1, 61–64. doi: 10.1128/jcm.1.1.61-64.1975

Rosenberg, S. M. (2001). Evolving responsively: adaptive mutation. *Nat. Rev. Genet.* 2, 504–515. doi: 10.1038/35080556

Rossi, E., Paroni, M., and Landini, P. (2018). Biofilm and motility in response to environmental and host-related signals in gram negative opportunistic pathogens. *J. Appl. Microbiol.* 125, 1587–1602. doi: 10.1111/jam.14089

Ryan, M. P., and Adley, C. C. (2013). The antibiotic susceptibility of water-based bacteria *Ralstonia pickettii* and *Ralstonia insidiosa*. *J. Med. Microbiol.* 62, 1025–1031. doi: 10.1099/jmm.0.054759-0

Ryan, M. P., Pembroke, J. T., and Adley, C. C. (2006). *Ralstonia pickettii*: a persistent gram-negative nosocomial infectious organism. *J. Hosp. Infect.* 62, 278–284. doi: 10.1016/j.jhin.2005.08.015

Ryan, M. P., Pembroke, J. T., and Adley, C. C. (2011). Genotypic and phenotypic diversity of *Ralstonia pickettii* and *Ralstonia insidiosa* isolates from clinical and environmental sources including high-purity water. Diversity in *Ralstonia pickettii*. *BMC Microbiol.* 11:194. doi: 10.1186/1471-2180-11-194

Singh, N. K., Wood, J. M., Karouia, F., and Venkateswaran, K. (2018). Succession and persistence of microbial communities and antimicrobial resistance genes associated with international Space Station environmental surfaces. *Microbiome.* 6:204. doi: 10.1186/s40168-018-0585-2

Sriswasdi, S., Yang, C.-C., and Iwasaki, W. (2017). Generalist species drive microbial dispersion and evolution. *Nat. Commun.* 8:1162. doi: 10.1038/s41467-017-01265-1

Stelzmueller, I., Biebl, M., Wiesmayr, S., Eller, M., Hoeller, E., Fille, M., et al. (2006). *Ralstonia pickettii*-innocent bystander or a potential threat? *Clin. Microbiol. Infect.* 12, 99–101. doi: 10.1111/j.1469-0691.2005.01309.x

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J. R., Easson, C., Astudillo-García, C., et al. (2016). Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* 7:11870. doi: 10.1038/ncomms11870

Thornhill, S. G., and Kumar, M. (2018). Biological filters and their use in potable water filtration systems in spaceflight conditions. *Life Sci Space Res* 17, 40–43. doi: 10.1016/j.lssr.2018.03.003

Tolker-Nielsen, T. (2015). Biofilm development. *Microbiol Spectr.* 3:2014. doi: 10.1128/microbiolspec.MB-0001-2014

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W., and Corander, J. (2018). RhierBAPS: an R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res* 3:93. doi: 10.12688/wellcomeopenres.14694.1

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., et al. (2020). PHI-base: the pathogen-host interactions database. *Nucleic Acids Res.* 48, D613–D620. doi: 10.1093/nar/gkz904

Vaishampayan, A., de Jong, A., Wight, D. J., Kok, J., and Grohmann, E. (2018). A novel antimicrobial coating represses biofilm and virulence-related genes in methicillin-resistant *Staphylococcus aureus*. *Front. Microbiol.* 9:221. doi: 10.3389/fmicb.2018.00221

Vaishampayan, A., and Grohmann, E. (2019). Multi-resistant biofilm-forming pathogens on the international Space Station. *J. Biosci.* 44:125. doi: 10.1007/s12038-019-9929-8

Wei, W., Ho, W. C., Behringer, M. G., Miller, S. F., Bcharah, G., and Lynch, M. (2022). Rapid evolution of mutation rate and spectrum in response to environmental and population-genetic challenges. *Nat. Commun.* 13:4752. doi: 10.1038/s41467-022-32353-6

Wellenreuther, M., Mérot, C., Berdan, E., and Bernatchez, L. (2019). Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* 28, 1203–1209. doi: 10.1111/mec.15066

Wong, W. C., Dudinsky, L. A., Garcia, V. M., Ott, C. M., and Castro, V. A. (2010). Efficacy of various chemical disinfectants on biofilms formed in spacecraft potable water system components. *Biofouling* 26, 583–586. doi: 10.1080/08927014.2010.495772

Xu, Z., and Yuan, C. (2022). Molecular epidemiology of *Staphylococcus aureus* in China reveals the key gene features involved in epidemic transmission and adaptive evolution. *Microbiol Spectr.* 10:e0156422. doi: 10.1128/spectrum.01564-22

Yabuuchi, E., Kosako, Y., Yano, I., Hotta, H., and Nishiuchi, Y. (1995). Transfer of two *Burkholderia* and an *Alcaligenes* species to *Ralstonia* gen. Nov.: proposal of *Ralstonia pickettii* (Ralston, Palleroni and Doudoroff 1973) comb. Nov., *Ralstonia solanacearum* (Smith 1896) comb. Nov. and *Ralstonia eutropha* (Davis 1969) comb. Nov. *Microbiol. Immunol.* 39, 897–904. doi: 10.1111/j.1348-0421.1995.tb03275.x

Yin, Z., Zhang, S., Wei, Y., Wang, M., Ma, S., Yang, S., et al. (2020). Horizontal gene transfer clarifies taxonomic confusion and promotes the genetic diversity and pathogenicity of *Plesiomonas shigelloides*. *mSystems* 5, e00448–e00420. doi: 10.1128/mSystems.00448-20

Yuan, C., Wei, Y., Zhang, S., Cheng, J., Cheng, X., Qian, C., et al. (2020). Comparative genomic analysis reveals genetic mechanisms of the variety of pathogenicity, antibiotic resistance, and environmental adaptation of *Providencia* genus. *Front. Microbiol.* 11:572642. doi: 10.3389/fmicb.2020.572642

Yuan, C., Yin, Z., Wang, J., Qian, C., Wei, Y., Zhang, S., et al. (2019). Comparative GenomicAnalysis of Citrobacter and key genes essential for the pathogenicity of *Citrobacter koseri*. *Front. Microbiol.* 10:2774. doi: 10.3389/fmicb.2019.02774

Zhu, Q., Kosoy, M., and Dittmar, K. (2014). HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics* 15:717. doi: 10.1186/1471-2164-15-717

TYPE Original Research
PUBLISHED 14 March 2024
DOI 10.3389/fmicb.2024.1335081

OPEN ACCESS

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Chien-Jui Huang,
National Chiayi University, Taiwan
Dharmendra Kumar Soni,
Uniformed Services University of the Health Sciences, United States

*CORRESPONDENCE
Wen Yu
✉ yuwen_1322@163.com
Ruimei Geng
✉ gengruimei@caas.cn

RECEIVED 08 November 2023
ACCEPTED 01 February 2024
PUBLISHED 14 March 2024

CITATION
Xiao Z, Li G, Yang A, Liu Z, Ren M, Cheng L, Liu D, Jiang C, Wen L, Wu S, Cheng Y, Yu W and Geng R (2024) Comprehensive genome sequence analysis of *Ralstonia solanacearum* gd-2, a phylotype I sequevar 15 strain collected from a tobacco bacterial phytopathogen.
Front. Microbiol. 15:1335081.
doi: 10.3389/fmicb.2024.1335081

Comprehensive genome sequence analysis of *Ralstonia solanacearum* gd-2, a phylotype I sequevar 15 strain collected from a tobacco bacterial phytopathogen

Zhiliang Xiao[1], Guangcan Li[1,2], Aiguo Yang[1], Zhengwen Liu[1], Min Ren[1], Lirui Cheng[1], Dan Liu[1], Caihong Jiang[1], Liuying Wen[1], Shengxin Wu[3], Yazhi Cheng[3], Wen Yu[3]* and Ruimei Geng[1]*

[1]The Key Laboratory for Tobacco Gene Resources, Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China, [2]Qingdao Agricultural University, College of Agriculture, Qingdao, China, [3]Fujian Institute of Tobacco Agricultural Sciences, Fuzhou, China

Introduction: Plant bacterial wilt is an important worldwide disease caused by *Ralstonia solanacearum* which is a complex of species.

Methods: In this study, we identified and sequenced the genome of *R. solanacearum* strain gd-2 isolated from tobacco.

Results: Strain gd-2 was identified as *R. solanacearum* species complex (RSSC) phylotype I sequevar 15 and exhibited strong pathogenicity to tobacco. The genome size of gd-2 was 5.93 Mb, including the chromosomes (3.83 Mb) and the megaplasmid (2.10 Mb). Gene prediction results showed that 3,434 and 1,640 genes were identified in the chromosomes and plasmids, respectively. Comparative genomic analysis showed that gd-2 exhibited high conservation with ten highly similar strain genomes and the differences between gd-2 and other genomes were mainly located at positions GI12-GI14. 72 type III effectors (T3Es) were identified and RipAZ2 was a T3E specific to gd-2 compared with other eight sequenced strain.

Discussion: Our study provides a new basis and evidence for studying the pathogenic mechanism of *R. solanacearum*.

KEYWORDS
***Ralstonia solanacearum*, type III effectors, whole-genome sequencing analysis, virulence factors, comparative genomic analysis**

1 Introduction

Plant bacterial wilt is an important worldwide disease caused by *Ralstonia solanacearum*, which can be transmitted through soil, irrigation, plants, and seed potatoes (Genin and Denny, 2012). This pathogen has a wide host range and can infect more than 200 plant species belonging to more than 50 families, including monocotyledons and dicotyledons, such as potatoes, tomatoes, eggplants, peanuts, tobacco, bananas, and ginger (Paret et al., 2010; Qian et al., 2012; Yuliar et al., 2015; Sharma, 2021; Huang et al., 2023). Bacterial wilt is widely distributed in tropical, subtropical, and temperate regions (Elphinstone et al., 2005). Peanut wilt generally causes a 10–20% reduction in production, and, production in severe cases was reduced up to 50% or even halted (Chen et al., 2020). Ginger wilt caused by *R. solanacearum* remains the biggest obstacle to ginger production (Mao et al., 2017). In 1880, Bacterial wilt in tobacco was first discovered in Granville, USA, then it was classified as an important disease in tobacco production subsequently because of the enormous potential threat posed to the tobacco industry (Hayward, 2003). And bacterial wilt in tobacco has subsequently spread throughout the world, including in the United States, Indonesia, Japan, South Korea, and Australia (Prokchorchik et al., 2020).

Ralstonia solanacearum has complex physiological and biochemical characteristics. It is a gram-negative rod-shaped bacterium with an optimum growth temperature of approximately 32°C and a pH of 6.6. When cultured on TTC medium, *R. solanacearum* generally exhibits a central reddish color surrounded by a milky white irregular shape and exhibits strong fluidity under high light conditions (Kang et al., 2008). There are two internationally recognized traditional taxonomic methods for *R. solanacearum*. One method divides *R. solanacearum* into five physiological variants based on their host range (Buddenhagen et al., 1962). The other method divides *R. solanacearum* into five biochemical variants based on their utilization of carbohydrates (lactose, maltose, cellobiose, mannitol, sorbitol, and xylitol) (He, 1983). According to the diversity of *R. solanacearum* in different hosts and different geographical origins, Fegan and Prior proposed a new evolutionary taxonomic framework based on the analysis of the 16S-23S rDNA gene spacer region sequence *endoglucanase* (*egl*) gene and *hypersensitive response and pathogenicity* (*hrpB*) gene, which reflects the genetic evolution and geographical origins of *R. solanacearum* better. Its evolutionary taxonomic framework includes four different levels of taxonomic units: species, phylotype, sequevar and clone (Fegan and Prior, 2005). These evolutionary types reflect their different geographical origins: Asian (phylotype I), American (phylotype IIA and phylotype IIB), African (phylotype III), and Indonesian (phylotype IV) (Castillo and Greenberg, 2007). Each evolutionary type can be further subdivided into different sequence types (sequevars), and different sequence types may contain different strains with similar pathogenicity or consistent geographical origins. According to the homology of the *egl* gene sequence in the strains, each evolutionary type of strain is divided into multiple different sequence variants; 55 sequence variants have been identified to date (Liu et al., 2017; Greenrod et al., 2023).

The pathogenesis and regulatory process of *R. solanacearum* are very complex. The main virulence factors include the type I, II, III, IV, V, VI secretion system (T1SS, T2SS, T3SS, T4SS, T5SS, T6SS), extracellular polysaccharides (EPSs) and extracellular proteins (EXPs). Among them, EPSs play a crucial role in pathogenicity of bacterial (Kang et al., 2002; Valls et al., 2006; Tsai et al., 2019). *R. solanacearum* can spread through soil, and it enters the plant roots and invades the vascular bundles of the plant and rapidly spreads to the aboveground tissues through the vascular bundle system. The typical symptoms of diseases caused by *R. solanacearum* infection are browning of the xylem, preferential growth of leaves, and plant wilting. After entering the host, *R. solanacearum* secretes more than 30 effector proteins through the type II secretion system (T2SS), including various cell wall-degrading enzymes. The most studied effector proteins are pectinolytic enzymes and cellulose hydrolytic enzymes, which play an important role in the colonization of *R. solanacearum* (Liu et al., 2005; Tsujimoto et al., 2008; Sharma et al., 2020). The T3SS plays an important role in the interaction between *R. solanacearum* and its host (Alfano and Collmer, 2004; Coll and Valls, 2013). All the type III effectors (T3Es) of *R. solanacearum* are located on the large plasmid of the bacterium, known as the *hrp*, which is approximately 23–30 kb. When this region is mutated, the host cannot exhibit a hypersensitive response or cause plant disease (Mukaihara et al., 2010; Ran et al., 2014). The effector proteins of *R. solanacearum* exhibit widespread gene-level transfer and significant intraspecies genetic differentiation. Peeters et al. (2013a,b) sorted the effector proteins of *R. solanacearum* and unified their nomenclature based on their genetic relationships using the general term Rip to name all T3E genes, obtaining 94 Rip genes and 16 candidate T3E genes. Sabbagh et al. (2019) updated the database published by Peeters et al. (2013a,b) and generated a pangenomic library containing 102 T3Es and 16 hypothetical T3Es. The functions of the effector proteins of *R. solanacearum* include interfering with the basic defense response of plants, interfering with host plant metabolic processes, promoting infection, and stimulating host plant immune responses (Tasset et al., 2010; Landry et al., 2020; Cheng et al., 2021; Schachterle and Huang, 2021).

In recent years, the completion of whole-genome sequencing of *R. solanacearum* has laid the foundation for researchers to elucidate the molecular mechanism of disease pathogenesis at the genomic level. The genome of *R. solanacearum* is approximately 5.8 Mb, dominated by two circular replicons. Housekeeping genes and some virulence genes are located on the chromosomes, while important virulence factors, such as T3SS and EPS which determine the pathogenicity of *R. solanacearum*, are located on megaplasmids. Salanoubat et al. (2002) isolated the *R. solanacearum* strain GMI1000 from tomato plants and completed genome sequencing using this strain as a material, which allowed for further research on its pathogenesis mechanism and identification of plant resistance improvement strategies. Currently, the gene data of the strain can be obtained on three database platforms, NCBI,[1] Ralsto T3E[2] and *R. solanacearum* sp.,[3] which play an important role in analyzing the diversity and evolution of the *R. solanacearum* genome, studying the genes affecting host range, and determining the comprehensive regulatory mechanism controlling bacterial virulence (Guidot et al., 2009; Peeters et al., 2013a,b; Tan et al., 2019). Genomic islands (GI) are an important form of horizontal gene transfer (HGT), which contain genes related to various biological functions. The genes carried by GI can often bring selective advantages to bacteria. According to the different genes contained, GI can be generally divided into virulence islands, drug resistance islands, metabolic islands, symbiotic islands (Shrivastava et al., 2010). Then gene islands, secreted proteins are generally considered when identifying virulence factors, which play a key role in enhancing the pathogenic efficacy of pathogens (Stritzler et al., 2018; Choi et al., 2020). The NCBI database has published the complete draft genome of 145 *R. solanacearum*. These Ralstonia strains were mainly isolated from tomato (GMI1000, FJAT-1458), eggplant (EP1, RS-09-161), pepper (RS-10-244, KACC10709), tobacco (CQPS-1, FQY-4), potato (UY031), sesame (SEPPX05) and plantain plants (UW163) (Salanoubat et al., 2002; Ahn et al., 2011; Cai et al., 2015; Asolkar and Ramesh, 2018). The genomic

[1] https://www.ncbi.nlm.nih.gov/genome/genomes/490

[2] https://iant.toulouse.inra.fr/bacteria/annotation/site/prj/T3Ev3/

[3] http://sequence.toulouse.inra.fr/R.solanacearum

data of Ralstonia isolated from tobacco include Y45 (phylotype I, sequevar 17), FQY-4 (phylotype I, sequevar 17), CQPS-1 (phylotype I, sequevar 17), FJ1003 (phylotype I, sequevar 14), and SL1931 (race 1, biovar 3 strain) (Cao et al., 2013; Liu et al., 2017). Using the genomics of Ralstonia to understand and explore the regulatory mechanism of virulence differentiation and host adaptation will provide an important theoretical basis for targeted prevention and control of Ralstonia.

In this study, we report the isolation of the *R. solanacearum* strain gd-2 from tobacco plants in Fujian Province, China. This strain belongs to phylotype I sequence 15 and exhibits strong pathogenicity to tobacco. We performed whole-genome sequencing and assembly to obtain the genome framework of gd-2. In addition, we performed functional annotation of the gd-2 genome and compared it with other published *R. solanacearum* genome sequences using comparative genomics analysis to explore whether gd-2 has genome segments or genes related to host specificity, providing new evidence for ultimately analyzing the pathogenic specificity of *R. solanacearum* and the prevention and control of bacterial wilt.

2 Materials and methods

2.1 Strain gd-2 classification and pathogenicity identification

The bacterial strain gd-2 was isolated from Fujian Province, China, and is preserved by the Tobacco Research Institute of the Chinese Academy of Agricultural Sciences. A bacterial genomic DNA extraction kit (TIANGEN, Beijing, China) was used to extract the genomic DNA of the strain. Polymerase chain reaction (PCR) amplification was performed using a composite PCR of the phylotype type of *R. solanacearum*, and the primers are shown in Table 1. The band information was observed through the gel imager, the tested strain was determined to belong to *R. solanacearum* according to whether there were 759/760 bands, and the evolutionary type of the strain was determined by the size of the band.

Three tobacco cultivars hongda, CHB and K326 were used for pathogenicity test of tobacco bacterial wilt for gd-2, which were common resistance and susceptible control variety (Li et al., 2015; Cao et al., 2013; Pan et al., 2021). And hongda existed high susceptible (HS) to susceptible (S) for bacterial wilt, CHB existed S and K326 existed middle resistance (MR) to resistance (R), respectively. After being activated for 36 h by inoculation and streaking on a TTC culture medium plate, single colonies were picked up with disposable inoculation rings and inoculated into NB liquid medium. The inoculated medium was placed in a shaker at 28°C and 220 r/min for 24 h, yielding a bacterial solution with a concentration of approximately 1.0×10^9 CFU/mL (OD_{600} of 1.0). The bacterial solution was diluted with deionized water to 1.5×10^8 CFU/mL. When the tobacco seedlings reached the five-leaf stage, 30 mL of diluted bacterial suspension was introduced into the bottom of each seedling at a standard rate, keeping the temperature of the bacterial solution at approximately 30°C and the humidity at approximately 80%. Twenty seedlings were inoculated for each variety, and the incidence of bacterial wilt was investigated at 3, 6, 9, 12, 15, and 18 days after inoculation.

2.2 Genomic sequencing and assembly

The third-generation sequencing technology platform PacBio Sequel II sequencer (Pacific Biosciences[4]) was used for genome sequencing, which was sequenced by Shanghai Winnerbio Technology Co., Ltd. (Shanghai, China) using Illumina NovaSeq 6000. The sequencing strategy was single-molecule real-time (SMRT) sequencing. Subsequently, the genome sequencing data were analyzed by GC depth analysis and K-mer frequency distribution analysis to determine whether there was contamination from other species or large fragments from other sources. The third-generation PacBio Sequel II platform and the second-generation sequencing platform Illumina NovaSeq 6000 were used to construct large fragment libraries (10–20 kb) and small fragment libraries (~400 bp) from DNA samples that passed quality control. The raw data were obtained by sequencing on different platforms. Canu V2.2[5] was used to assemble the third-generation data independently. Then, the final assembly result was corrected by using Pilon V1.24[6] and the second-generation data to obtain the final assembly of the bacterial genome.

2.3 Gene prediction and analysis

The assembled genome sequence was used to predict the coding sequence (CDS) of coding genes using Glimmer3[7] software, transfer RNA (tRNA) genes using tRNAscan-SE[8] software, and ribosomal RNA (rRNA) genes using Barrnap[9] software. Gene function annotation was performed by comparing the protein sequence file of the gene with the database. The relevant databases included the Nonredundant Protein Database (NR), Swiss-Prot database,[10] Pfam database,[11] Gene Ontology (GO) database,[12] and Kyoto Encyclopedia of Genes and Genomes (KEGG) database.[13] DIOMAN[14] was used for sequence alignment analysis.

Circos Version 0.69-6[15] software was used to draw genome circles for the obtained chromosomes and plasmids. The default scanning map ordered all scaffolds from large to small or from small to large and concatenated them into a sequence to draw circles without distinguishing direction. The default information from the outer to the inner circle corresponded to genome size identification, positive-and negative-strand gene information, noncoding RNA (ncRNA), GC content, and GC skew. In addition, IslandViewer 4[16] was used for genome island prediction using IslandPath-DIMOB, Islander, and other methods. Prophage prediction was carried out using

[4] http://www.pacb.com
[5] https://github.com/marbl/canu
[6] https://github.com/broadinstitute/pilon
[7] http://ccb.jhu.edu/software/glimmer/index.shtml
[8] http://trna.ucsc.edu/software/
[9] https://github.com/tseemann/barrnap/
[10] https://web.expasy.org/docs/swiss-prot_guideline.html
[11] http://pfam.xfam.org/
[12] https://geneontology.org/
[13] https://www.kegg.jp/kegg/
[14] http://github.com/bbuchfink/diamond
[15] http://www.circos.ca
[16] https://www.pathogenomics.sfu.ca/islandviewer/

TABLE 1 Comparison of core type III effectors (T3Es) genes of strain gd-2 with strain CFBP2957, CMR15, CQPS-1, FQY_4, GMI1000, Po82, PSI07, Y45.

T3E_ Name	GD-2 Gene ID	CFBP2957	CMR15	CQPS-1	FQY_4	GMI1000	Po82	PSI07	Y45
RipA1	gene1265	absent	absent	absent	100/98	100/97	absent	absent	100/98
RipA2	pA_gene0050	99/78	100/92	100/98	100/99	100/99	96/78	100/77	100/99
RipA3	pA_gene0823	absent	100/89	100/99	100/98	100/98	99/70	99/81	100/99
RipA4	pA_gene0822	100/75	100/90	100/99	100/98	100/99	100/76	100/80	100/99
RipA5	pA_gene1041	98/78	99/93	100/99	100/99	93/99	99/82	99/81	100/99
RipAA	gene2855	68/74	69/76	absent	99/99	absent	67/77	76/79	99/99
RipAB	pA_gene0793	100/72	100/93	100/100	100/100	100/99	100/70	100/75	100/100
RipAC	pA_gene0794	absent	absent	96/100	100/99	100/99	absent	100/73	96/100
RipAD	pA_gene1572	absent	56/79	absent	66/96	53/97	absent	52/71	53/96
RipAE	gene3155	absent	absent	100/98	100/98	100/98	absent	99/76	100/99
RipAF1	pA_gene0847	absent	absent	100/98	100/96	82/97	absent	absent	100/99
RipAI	pA_gene0831	99/81	100/95	100/100	100/99	62/100	92/81	100/84	absent
RipAJ	gene1300	100/70	100/71	absent	100/98	92/99	absent	absent	100/99
RipAK	gene1017	absent	absent	92/99	92/98	84/99	absent	absent	88/99
RipAL	pA_gene0942	82/80	99/83	84/100	98/100	absent	82/80	100/98	94/100
RipAM	gene0169	89/73	100/93	absent	100/100	100/100	89/73	100/83	100/100
RipAN	pA_gene0824	99/70	100/84	100/98	100/98	100/97	98/70	98/73	100/98
RipAO	pA_gene0790	76/74	100/84	100/98	82/100	98/97	72/73	absent	absent
RipAP	pA_gene1238	100/77	100/81	100/99	absent	100/100	100/78	absent	100/99
RipAQ	pA_gene0784	absent	89/77	100/98	100/99	100/98	absent	absent	100/98
RipAR	pA_gene1256	absent	97/74	100/99	100/98	100/99	absent	57/73	100/99
RipAS	pA_gene1370	99/81	100/88	100/99	99/99	100/98	absent	absent	100/99
RipAU	pA_gene1443	82/74	83/80	83/99	83/100	83/99	58/70	absent	100/100
RipAV	pA_gene0934	99/86	100/87	100/97	100/99	100/98	83/86	absent	100/98
RipAW	pA_gene1459	absent	100/84	100/98	100/99	100/95	absent	absent	100/99
RipAX2	pA_gene0569	absent	absent	absent	absent	absent	absent	absent	absent
RipAY	pA_gene1039	absent	98/86	absent	100/97	100/96	absent	absent	100/99
RipAZ1	pA_gene1559	98/81	absent	100/99	100/99	100/99	95/83	97/91	100/99
RipAZ2	gene2616	absent	absent	absent	absent	absent	absent	absent	absent
RipB	gene3236	absent	absent	absent	absent	absent	absent	absent	59/100
RipBD	gene2483	absent	absent	100/100	absent	absent	63/94	absent	84/100
RipC1	pA_gene1259	absent	absent	absent	100/99	100/99	98/70	100/95	100/99
RipC2	pA_gene0585	absent	absent	absent	100/100	100/98	absent	absent	100/100
RipC2	gene2535	92/81	absent	100/83	97/85	97/82	absent	absent	96/75
RipC2	pA_gene0586	absent	absent	absent	100/98	79/99	absent	absent	absent
RipD	pA_gene0241	absent	absent	100/91	100/98	100/98	absent	100/82	100/99
RipE1	gene0066	100/85	96/87	100/99	100/99	100/98	94/85	100/79	94/99
RipE2	gene2507	100/88	absent	absent	absent	absent	100/89	100/98	100/99
RipF1_1	pA_gene1532	100/78	100/94	100/99	100/99	100/99	93/83	100/77	99/99
RipF1_2	pA_gene0754	100/81	100/89	100/81	100/81	100/98	93/76	100/83	100/99
RipG1	pA_gene0742	absent	absent	99/99	100/99	100/99	absent	absent	95/99
RipG2	pA_gene0991	100/72	99/80	100/97	100/98	100/98	100/72	absent	100/97
RipG3	pA_gene0027	absent	86/84	86/99	100/95	86/95	absent	absent	79/95

(Continued)

TABLE 1 (Continued)

T3E_ Name	GD-2 Gene ID	CFBP2957	CMR15	CQPS-1	FQY_4	GMI1000	Po82	PSI07	Y45
RipG4	gene1804	absent	100/80	absent	100/98	100/99	absent	absent	100/99
RipG5	gene1805	absent	99/87	100/98	97/98	97/98	absent	97/73	100/99
RipG6	gene2000	99/70	99/85	99/97	99/97	99/78	98/70	98/71	100/97
RipG7	gene1999	absent	absent	absent	absent	absent	absent	99/72	100/99
RipH1	gene1971	absent	99/82	100/97	100/98	100/96	absent	absent	100/99
RipH2	pA_gene0161	absent	98/82	100/97	100/98	100/96	absent	absent	absent
RipH3	pA_gene0104	100/75	100/86	100/95	100/95	100/95	96/78	98/70	100/95
RipI	gene3456	99/85	100/87	100/98	100/98	100/99	99/82	100/93	100/99
RipJ	gene1270	absent	absent	100/94	100/95	69/97	absent	absent	69/97
RipL	pA_gene0139	absent	99/81	100/97	100/97	99/97	absent	absent	96/98
RipM	gene1879	98/72	99/88	absent	100/98	100/99	99/82	99/80	100/98
RipN	pA_gene1159	99/75	100/89	100/98	100/98	100/98	99/74	99/73	100/97
RipO1	pA_gene0263	85/85	100/91	100/100	95/99	88/99	85/86	absent	100/98
RipQ	pA_gene1300	100/80	absent	100/97	100/99	100/99	94/71	absent	100/99
RipR	pA_gene1305	99/83	99/91	absent	100/99	100/99	100/79	100/82	100/99
RipS1	gene0034	96/87	96/90	97/71	100/72	94/99	96/88	96/75	96/99
RipS2	pA_gene1356	97/73	99/91	absent	99/99	93/98	93/74	95/77	97/98
RipS3	pA_gene0726	99/75	99/91	100/99	100/98	96/98	99/75	99/77	100/99
RipS4	gene1840	99/72	96/82	100/99	100/99	98/98	98/71	absent	100/99
RipS5	pA_gene0237	absent	100/85	100/99	100/99	100/99	100/80	99/75	100/99
RipS6	gene1271	absent	absent	absent	100/99	95/98	absent	absent	99/99
RipS8	gene1843	absent	absent	99/98	99/98	absent	absent	98/89	95/98
RipTAL	gene1818	absent	absent	83/98	100/85	83/98	absent	99/74	89/99
RipTPS	pA_gene0935	100/79	100/95	100/99	94/100	100/99	100/78	94/77	100/100
RipU	pA_gene1235	65/73	absent	absent	100/98	absent	65/79	65/78	65/99
RipV1	gene2007	absent	100/83	100/99	100/98	100/99	absent	100/70	100/95
RipW	gene0660	100/76	100/89	100/99	100/99	100/98	100/76	100/81	100/99
RipX	pA_gene0792	absent	100/76	100/95	100/100	100/95	absent	94/77	100/95
RipZ	pA_gene1049	absent	87/83	absent	100/98	100/98	absent	91/71	100/98

PHASTER,[17] which predicted prophage regions in the chromosome sequence of *R. solanacearum* and analyzed their related genes. CRISPRFinder (Ibtissem et al., 2007) was used to identify all potential CRISPR sequences on the genome, showing their location, the base composition of repeats, and the base composition of spacers. Virulence proteins defined as virulence properties molecules with invasive and toxic properties produced by bacteria, viruses, fungi. They are mainly used to enter host tissue cells by suppressing or evading host immune responses when microorganisms infect hosts and obtain nutrients and self-proliferation from hosts. The Virulence Factors of Pathogenic Bacteria (VFDB) database[18] was used for virulence gene prediction. The screening thresholds used were a similarity of 80% or greater, a coverage of 60% or greater, and an E value of 1e-5. The comprehensive antibiotic resistance database (CARD[19]) was used for comparative analysis and screening of candidate drug resistance-related genes using screening thresholds of 80% similarity, 60% coverage, and an E value of 1e-5. Based on the Pathogen Host Interactions (PHI-base) database, DIAMOND software (screening threshold E value ≤1e-5) was used to align the amino acid sequence of the sequenced strain with the database. TMHMM (TMHMM Server V 2.0[20]) software was used to predict transmembrane proteins and their structural information among five common transmembrane structures: (a) type I transmembrane, (b) type II transmembrane, (c) multipass transmembrane, (d) lipid

[17] http://phaster.ca/

[18] http://www.mgc.ac.cn/VFs/main.htm

[19] http://arpcard.Mcmaster.ca

[20] http://www.cbs.dtu.dk/services/TMHMM-2.0/

chain-anchored membrane, and (e) GPI-anchored membrane. The SignalP[21] tool was used to predict the signal peptide region of each protein and then combined with the analysis results of transmembrane domains to select proteins with a signal peptide structure but without a transmembrane domain as candidate secreted proteins. The secondary metabolite synthesis gene clusters of the samples were predicted using the antiSMASH bacterial database.

2.4 Comparative genomic analysis

We conducted comparative genomic analysis via two methods to further investigate the structural characteristics and key genes of plasmids. One method identified plasmid-specific regions and mutation hotspots through comparative genomic circle diagrams, and the other method identified the structural differences of local gene clusters through gene cluster comparison. We conducted comparative analysis of the 10 reference genome sequences with the large plasmid in the gd-2 genome using BRIGV0.9.5 software[22] and constructed comparative genomic circle diagrams. In addition, we used EasyFigV2.2.3[23] to conduct detailed comparative gene cluster analysis of the target region.

2.5 Virulence genes analysis

In this study, we focused on analyzing the T3SS, T4SS, T6SS, as these secretion systems and effectors are often closely related to the pathogenic ability of bacteria. The annotation analysis was conducted to annotate T6SS using the software T6SS_finder with thresholds of identity ≥50% and E value $\leq 1e^{-5}$. The annotation analysis of the T3Es was conducted using BLAST+ with more than 80% identity and more than 60% gene alignment coverage. Based on the Type VI effector database summarized in the SecReT6 database, annotation analysis of the Type VI effector of the strain gd-2 was conducted using BLAST+ with over 80% identity and over 60% gene alignment coverage. Through HMMER3 software, various genes in the two-component signal transduction system in the genome were obtained and analyzed using the Pfam database combined with the structural domain characteristics of histidine kinases and response regulatory proteins. The genes were divided into three categories, regulator, sensor, and hybrid. We also analyzed the number of chemotaxis genes and recorded their detailed annotation information. The quorum sensing genes were analyzed to identify genes and pathways related to quorum sensing by comparison and analysis with the KEGG database.

2.6 Comparative analysis of T3Es in different *Ralstonia solanacearum* strains

The effector protein of the T3SS system in *R. solanacearum* is the main determinant protein of its pathogenicity. We identified the distribution and variation information of T3Es in gd-2 and nine other sequenced and published *R. solanacearum* genomes, including phylotype I strain GMI1000 (BioProject: PRJNA13); phylotype IIA strain CFBP2957 (BioProject: PRJNA224116); phylotype III strain CMR15 (BioProject: PRJEA50681); phylotype IIB strain Po82 (BioProject: PRJEA50683); phylotype IV strain PSI07 (BioProject: PRJNA66837); and three phylotype I *R. solanacearum* sequence variants that can infect tobacco, including sequevar 13 strain CQPS-1 (BioProject: PRJNA331070), sequevar 17 strain FQY_4 (BioProject: PRJNA182081), and sequevar 54 strain Y45 (BioProject: PRJNA224116). The localized T3E database to identify T3Es in nine published *R. solanacearum* genomes were constructed firstly. Using BLAST+, the blast stragegy were the E value 1e-^{5}, the over 60% coverage and the over 80%identity. And the common and unique T3Es among the nine *R. solanacearum* strains were compared using the Venn diagram to count the gene distribution and sequence variation of the T3Es in gd-2, the candidate T3E genes of the other eight strains were aligned using BLAST+. We statistically analyzed the functional genes related to the *hrp* gene cluster in all nine strains, extracted the protein sequence files of the related genes and performed comparative display the genomes of the main *R. solanacearum* strains.

3 Results

3.1 Identification and pathogenicity detection of *Ralstonia solanacearum* gd-2

The strain gd-2 cultured on TTC medium exhibited a central reddish color surrounded by a milky white irregular shape, and its mobility was visible under high light conditions (Figure 1A). Agarose *gel* electrophoresis showed that the *R. solanacearum* strain gd-2 exhibited 144 bp and 280 bp bands, indicating that the strain is *R. solanacearum* phylotype I (Figure 1B:line 3). Amplification of the *egl* gene of the *R. solanacearum* strain gd-2 using the *endoglucanase*-specific primers Endo-F/Endo-R resulted in an 800 bp band, and sequencing and alignment results showed that the strain belonged to sequence variant 15 (Supplementary Table S2). After inoculating different varieties of tobacco with gd-2, a typical symptom of bacterial wilt was observed: leaves gradually showed wilting symptoms, and the infection spread from the lower leaves to the upper leaves. The stems gradually decayed until the entire tobacco plant died. There were differences in resistance among the tobacco varieties (Figure 1C) and the disease index and resistance performance are similar to other study results of different strains (Li et al., 2015; Cao et al., 2013; Pan et al., 2021). These results indicated that gd-2 meets the characteristics of typical *R. solanacearum* and has pathogenicity to tobacco.

3.2 Sequencing, assembly and annotations of the gd-2 genome

Total 7,666,395,879 bp base reads were identified from sequencing data. The GC depth and K-mer frequency distribution results showed that there was no contamination of miscellaneous bacteria in the sequencing data. After assembling the sequencing data, 3,828,519 bp chromosomes and 2,098,962 bp plasmids were obtained. Gene prediction results showed that 3,434 and 1,640 genes were identified in the chromosomes and plasmids, respectively. The predicted

[21] http://www.cbs.dtu.dk/services/SignalP
[22] http://sourceforge.net/projects/brig/
[23] http://mjsull.github.io/Easyfig/

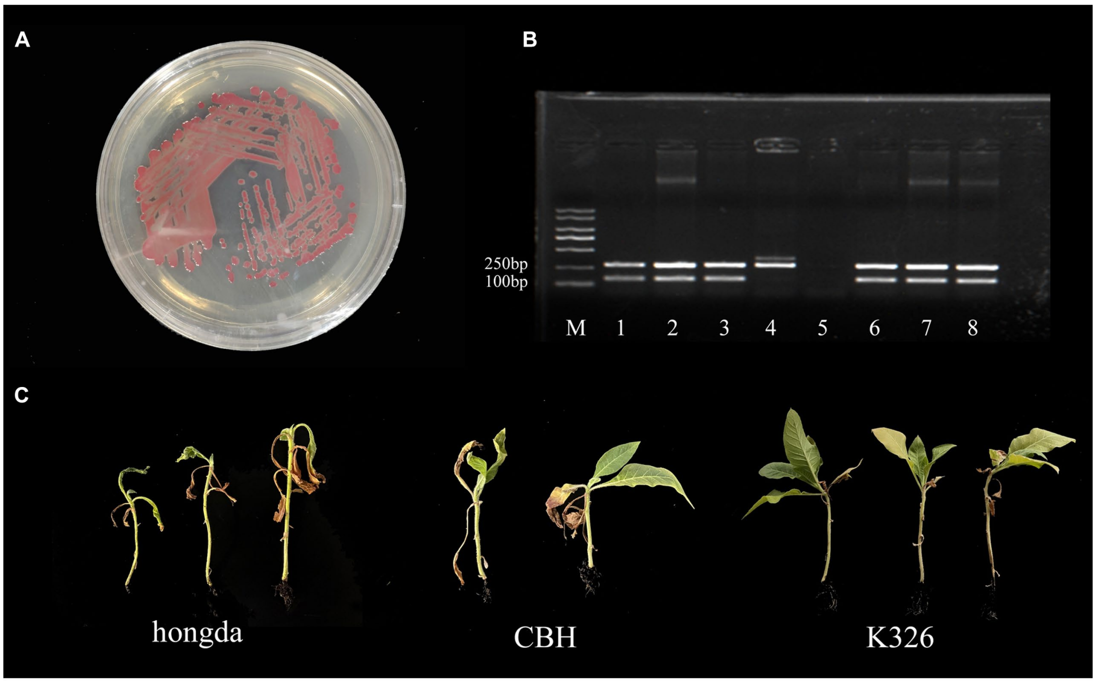


FIGURE 1
Phenotype, sequence variation, and infectivity identification of the gd-2 strain in tobacco. **(A)** Growth status of the gd-2 strain on TTC plates. **(B)** Electron microscopy observation of gd-2. Line M:marker 5000, line 1: Y45 (phylotype I, sequevar 17), line 2: FQY-4 (phylotype I, sequevar 17), Line 3: gd-2 (phylotype I, sequevar 15), Line 4: AM (phylotype II), Line 5: control, Line 6: HBES (phylotype I, sequevar 44), Line 7: SY-1 (phylotype I, sequevar 17), Line 8: SY-2 (phylotype I, sequevar 15). **(C)** Infectivity of gd-2 on different types of tobacco (hongda, CHB and K326).

noncoding RNAs included 59 tRNAs, 12 rRNAs, four 5S rRNAs, four 16S rRNAs, and four 23S rRNAs. Different databases identified 2,600–5,000 genes with functional annotations. The COG annotation classification, GO annotation classification, and KEGG annotation classification are shown in Figure 2. COG annotation classification involved 24 categories, and the five categories with the largest number of genes were amino acid transport and metabolism (417), transcription (393), general function prediction only (335), signal transduction mechanisms (299), and cell wall/membrane/envelope biogenesis (299); 172 genes had an unknown function. For GO annotation classification, the three biological process categories with the highest number of enriched genes were regulation of DNA-templated transcription, methylation, and phosphorylation; the three cellular component categories with the highest number of enriched genes were integral component of membrane, plasma membrane, and cytoplasm; and the three molecular function categories with the highest number of enriched genes were DNA binding, ATP binding, and metal ion binding. For KEGG annotation classification, global and overview maps, signal transduction, and membrane transport had the highest number of enriched genes.

The genome circle diagram include genome size identification, gene information on the positive and negative strands, ncRNA, GC content, GC skew, and other information (Figure 3). The genes carried by GIs usually confer selective advantages to bacteria. 17 GIs were identified, of which 11 originated from chromosomes and 6 from plasmids (Supplementary Table S3). Seven prophages were identified and of which six originated from chromosomes and one from plasmids (Supplementary Table S4). Three CRISPR_Cas systems were identified, of which one originated from chromosomes and two from plasmids (Supplementary Table S5). In addition, 130 genes were annotated as carbohydrate active enzyme-related genes (Supplementary Table S6), 68 genes were predicted as pathogen–host interaction-related genes (Supplementary Table S7), 1,172 genes were predicted to have transmembrane structures (Supplementary Table S8), 494 genes were predicted to be transporters (Supplementary Table S9), 848 genes contained signal peptide domains (Supplementary Table S10), and 705 genes were predicted to be secreted proteins (Supplementary Table S11).

3.3 Comparative genome analysis between the strain gd-2 and 10 highest similarity genomes

By BLAST alignment of gd-2 genome data with the NCBI database, 10 sequences with the highest similarity were identified, including six plasmid sequences of *R. solanacearum*, two chromosome sequences of *R. solanacearum*, and two plasmid sequences of *R. pseudosolanacearum* (Supplementary Table S12), the comparative genome circle diagram was shown in Figure 4A. The full length of gd-2-PlasmidA is 2,098,962 bp, with a GC content of 67.00%. Screening genes with identity >60% and coverage >90% predicted by CARD and VFDB, we identified nine possible antibiotic resistance genes and 18 candidated virulence factors. The antibiotic resistance genes are mainly related to various multidrug efflux pumps, such as *adeABC* gene *adeB*, *RosAB* gene *rosA* and *rosB*, *AcrAB-TolC* gene *acrB*, *AdeFGH* gene *adeF*, *MuxABC-OpmB* gene *MuxB*, *MdtABC-TolC* gene *MdtC* and *BaeR* which promotes the expression of *MdtABC* and *AcrD*

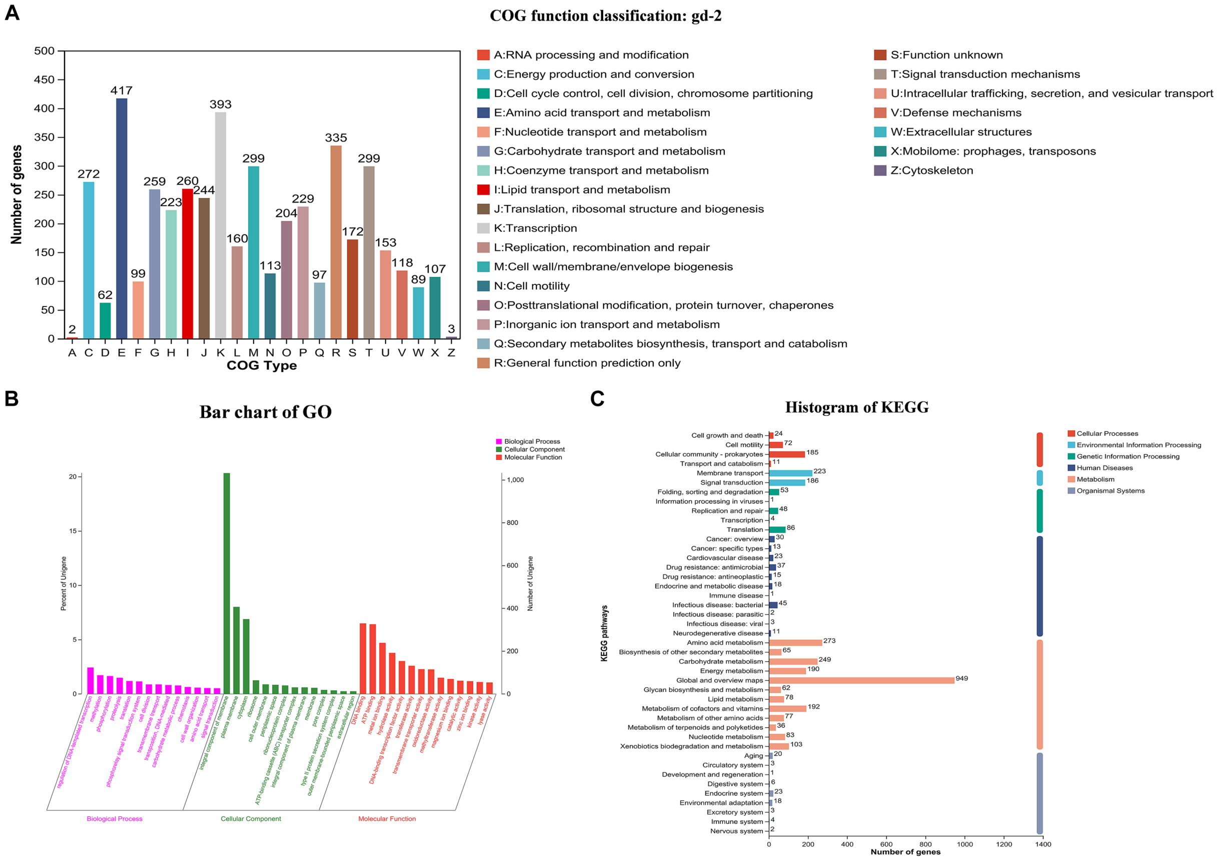


FIGURE 2
COG, GO and KEGG functional categories of protein-coding genes in the *R. solanacearum* gd-2 genome. **(A)** COG classification of gd-2. The horizontal axis is the 25 COG categories, and the vertical axis is the number of genes annotated to the relevant categories in the genome. **(B)** Clustering results of GO annotation. The horizontal axis is the 3 GO categories, and the vertical axis is the number of genes annotated to the relevant categories in the genome. **(C)** Histogram of KEGG analysis of gd-2. The horizontal axis is the number of genes, and the vertical axis is the number of genes annotated to each category in the genome.

efflux pumps. The virulence facors include *flagella* which encoding polar flagella needed for motility and macrophage invasion, *Cya* which encoding a dual-function toxin with adenylate cyclase and haemolytic activity, contribute as an anti-inflammatory protein and heat shock protein (Hsp) 60 which mediates complement-independent attachment to mammalian and amoebic host cells. In addition, five GIs, one prophage and two CRISPR elements were also identified on this plasmid. The differences between gd-2-PlasmidA and other genomes were mainly located at the position of GI12-GI14 (612,748–700,708 bp).

We performed a detailed comparative gene cluster analysis of the GI12-GI14 region of gd-2-PlasmidA with four highest similar strain B2 plasmid (GenBank: CP049788.1), strain R24 plasmid megaplasmid (GenBank: CP076122.1), strain 202 chromosome (GenBank: CP049789.1), and strain CQPS-1 chromosome (Figure 4B). The analysis showed that the GI12 sequence of gd-2-PlasmidA showed some differences in the two putative proteins before the second IS5 and parts of IS5 from other sequences except for the B2 plasmid. In contrast, the GI14 sequence showed high similarity among the various strains, although the six genes at the end of the chromosome were deleted in both strain 202 and strain CQPS-1. In addition, the prophage structure of the R24 plasmid megaplasmid showed the greatest difference from strain Ph07 of gd-2-PlasmidA, with the internal deletion of three consecutive *XerD* genes, which were inverted in the chromosome sequences of strain 202 and strain CQPS-1. Chromosome of strain 202 contained two copies of the same Ph07 strain-like sequence while chromosome strain CQPS-1 contained an additional copy of the Ph07 strain-like sequence, but it lacked many other genes. These results reflect that the mobile elements of GI12, Ph07 and GI14 may have undergone further gene loss or multiple gene copies after integration into the gd-2-PlasmidA-like plasmid of *R. solanacearum* resulting in significant differences in the plasmid.

3.4 Virulence genes of the strain gd-2

To further investigate the pathogenic mechanism of *R. solanacearum* gd-2, we predicted the structural genes of its various secretion systems. Among them, 66 secretion system structural genes were identified, including five T1SS, 23 T2SS, 10 T3SS, 16 T6SS, 11 Sec-SRP, and three twin-arginine protein translocation (Tat). The statistical analysis of the structural composition of the secretion system strains is shown in Figure 5. Among them, the T1SS secretion system includes *tolerant colicin* (*tolC*), *hemolysin D* (*hlyD*) and

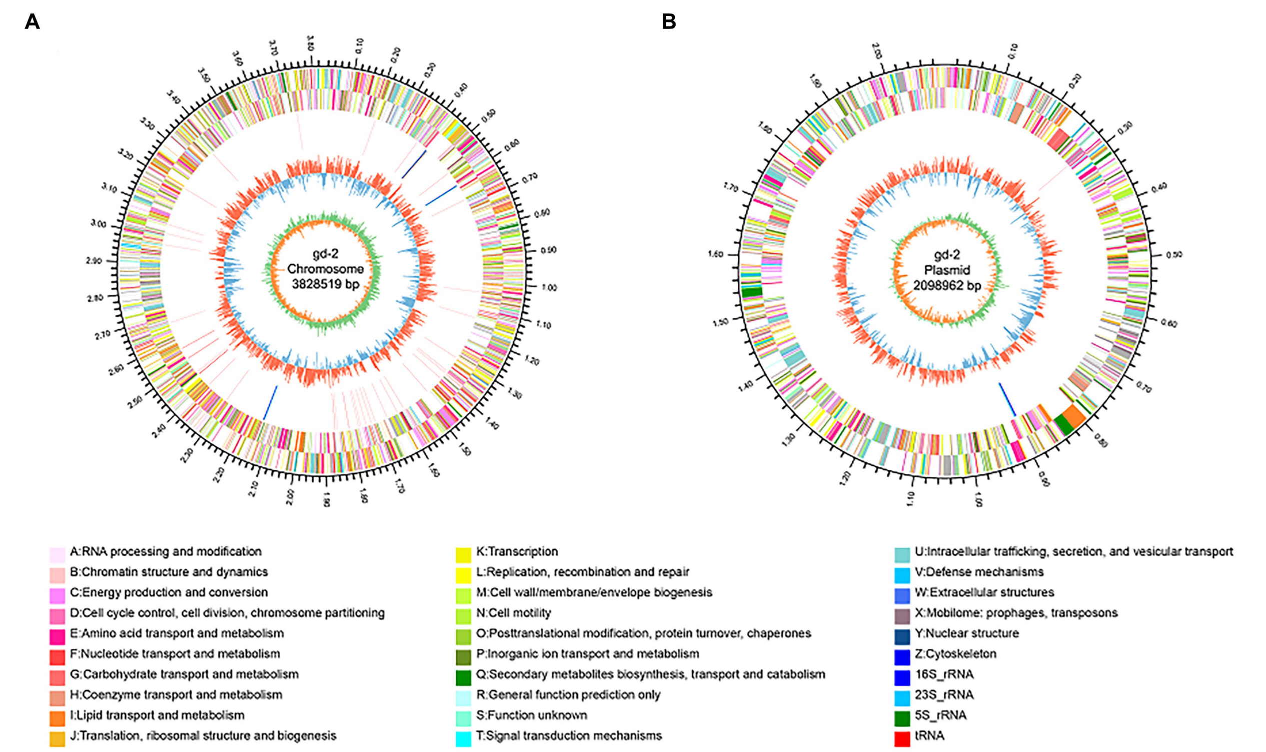

A
B
gd-2
Chromosome
3828519 bp
gd-2
Plasmid
2098962 bp
A:RNA processing and modification
B:Chromatin structure and dynamics
C:Energy production and conversion
D:Cell cycle control, cell division, chromosome partitioning
E:Amino acid transport and metabolism
F:Nucleotide transport and metabolism
G:Carbohydrate transport and metabolism
H:Coenzyme transport and metabolism
I:Lipid transport and metabolism
J:Translation, ribosomal structure and biogenesis
K:Transcription
L:Replication, recombination and repair
M:Cell wall/membrane/envelope biogenesis
N:Cell motility
O:Posttranslational modification, protein turnover, chaperones
P:Inorganic ion transport and metabolism
Q:Secondary metabolites biosynthesis, transport and catabolism
R:General function prediction only
S:Function unknown
T:Signal transduction mechanisms
U:Intracellular trafficking, secretion, and vesicular transport
V:Defense mechanisms
W:Extracellular structures
X:Mobilome: prophages, transposons
Y:Nuclear structure
Z:Cytoskeleton
16S_rRNA
23S_rRNA
5S_rRNA
tRNA

FIGURE 3
Genomic circle diagram of gd-2 (chromosome and plasmid). **(A)** Genomic circle diagram of gd-2. **(B)** Plasmid circle diagram of gd-2. The outermost circle of the circle diagram indicates the size of the genome; the second and third circles are the CDS on the positive and negative strands, with different colors indicating different functional classifications of COGs for the CDS; the fourth circle is rRNA and tRNA; the fifth circle is GC content, with outwards red portions indicating high GC content in the region. The higher the peak value is, the greater the difference from the average GC content. The inner blue part indicates that the GC content in this region is lower than the average GC content of the whole genome. The higher the peak value is, the greater the difference from the average GC content. The innermost circle is the GC skew value, which is calculated as G-C/G + C and can assist in determining the leading and lagging strands. Generally, the leading strand GC skew >0 and the lagging strand GC skew <0. It can also assist in determining the replication start point (minimum cumulative deviation) and end point (maximum cumulative deviation).

hlyB. The T2SS secretion system includes *gspF*, *gspE*, *gspD*, *gspM*, *gspL*, *gspK*, *gspJ*, *gspI*, *gspH*, *gspG*, and *gspC*. The T3SS secretion system located on plasmidA *includes yersinia* (*ysc*) gene *yscL*, *yscJ*, *yscU*, *yscV*, *yscQ*, *yscR*, and *yscS*. The T6SS secretion system includes *valine glycine repeat protein G* (*vgrG*), *intracellular multiplication protein L* (*impL*), *hcp*, *vasD*, and *impK*. The secretion system and signal recognition particle secretion system (Sec-SRP) includes *secA*, *secB*, *secD*, *secE*, *secF*, *secG*, *secY*, *ffh*, *yajC*, *ftsY*, and *yidC*. The Tat secretion system includes *tatA*, *tatB*, and *tatC*.

A total of 179 genes of the two-component signal transduction system were identified in the genome of gd-2, including 101 regulator genes, 60 sensor genes, and 18 hybrid genes (Supplementary Table S13). In addition, a total of 34 chemotaxis-related genes were predicted in gd-2 (Supplementary Table S14). 99 quorum sensing-related genes were predicted (Supplementary Table S15).

3.5 Comparative analysis of T3E analysis for gd-2 and other *Ralstonia solanacearum* strains

The nine strains of *R. solanacearum* were identified to have 54–75 candidate T3Es, with CQPS-1 (54) having the fewest T3Es and Po82 (75) having the most. Total 72 T3Es were identified from strain gd-2, which was comparable to the number of T3Es in GMI1000 (74), FQY-4 (70), and Y45 (69), all of which belong to phylotype I. Comparing the T3Es of gd-2 with different strains of *R. solanacearum*, 37 T3Es were shared by six strains, and the T3E unique to gd-2 was RipAZ2 (Figure 6A). The T3Es unique to strain PSI07 were RipA, RipBB, RipBF, RipE1_1, RipE1_2, RipG1_2, RipG1_3 and RipH4; the T3Es unique to Po82 were RipA5_1 and RipA5_2; and the T3Es unique to CMR15, GMI1000, and CFBP2957 were RipG8, RipAH, and RipK, respectively. Among gd-2, CQPS-1, Y45, and FQY-4, 45 shared T3Es were identified (Figure 6B). Each of the four strains contained a unique T3E, with RipAZ2 for gd-2, RipP2 for FQY_4, RipT for Y45, and RipBE for CQPS-1. The candidate T3Es comparison betwwen gd-2 and other eight strains showed that 17 T3Es were shared, including RipA2, RipA4, RipA5, RipAB, RipAN, RipE1, RipF1_1, RipF1_2, RipG6, RipH3, RipI, RipN, RipS1, RipS3, RipS4, RipTPS, and RipW (Table 1). Most T3Es had high similarity among different strains, but there were also cases of deletion, indicating that both conserved and specific T3Es exist in different strains.

The functional genes related to the *hrp* gene cluster in all nine samples were statistically analyzed. The number of functional genes related to the *hrp* gene cluster identified in different strains ranged from 26 to 30, with 30 *hrp* genes identified in gd-2 (Supplementary Table S15). The *hrp* genes in gd-2 were consistent

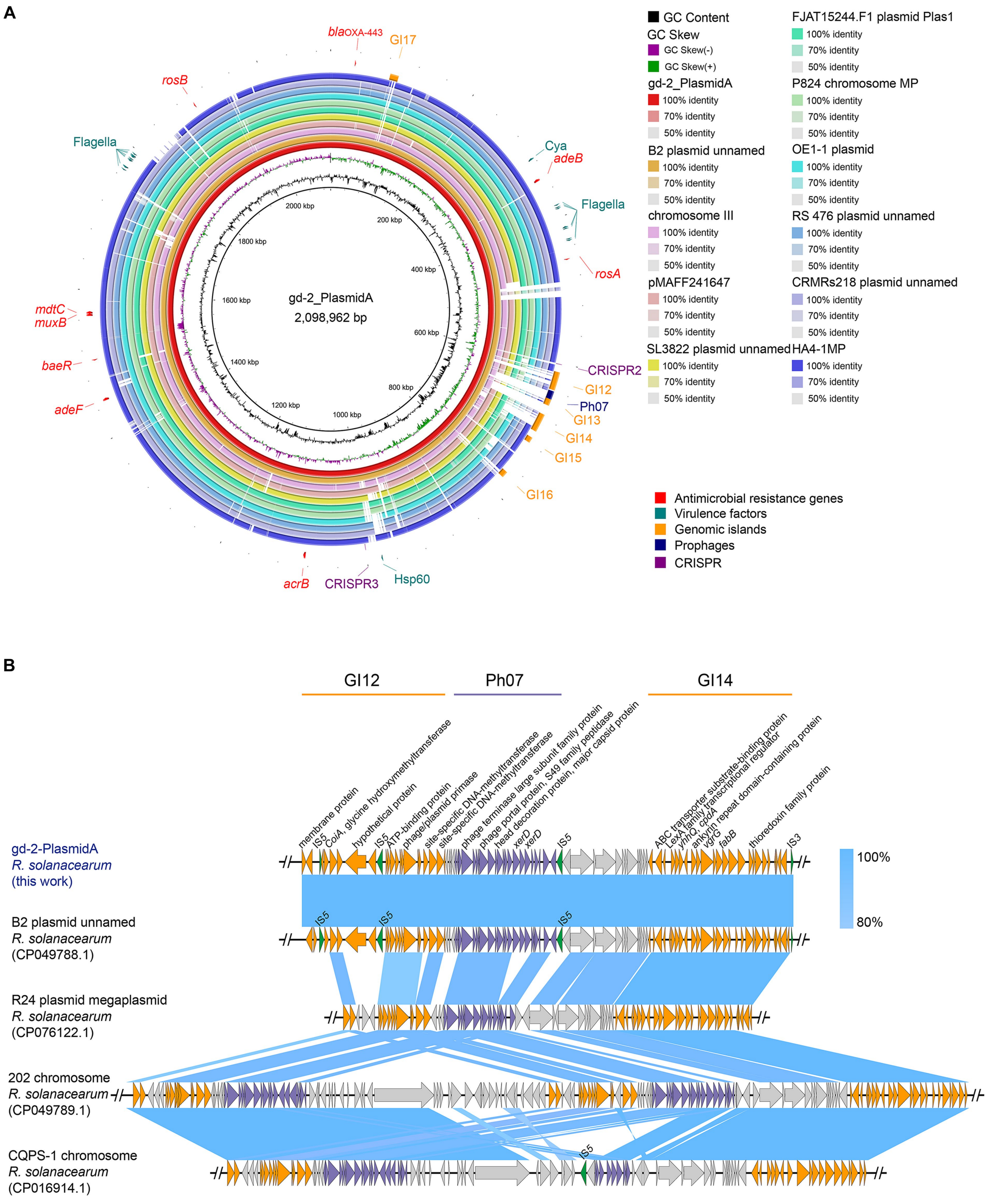
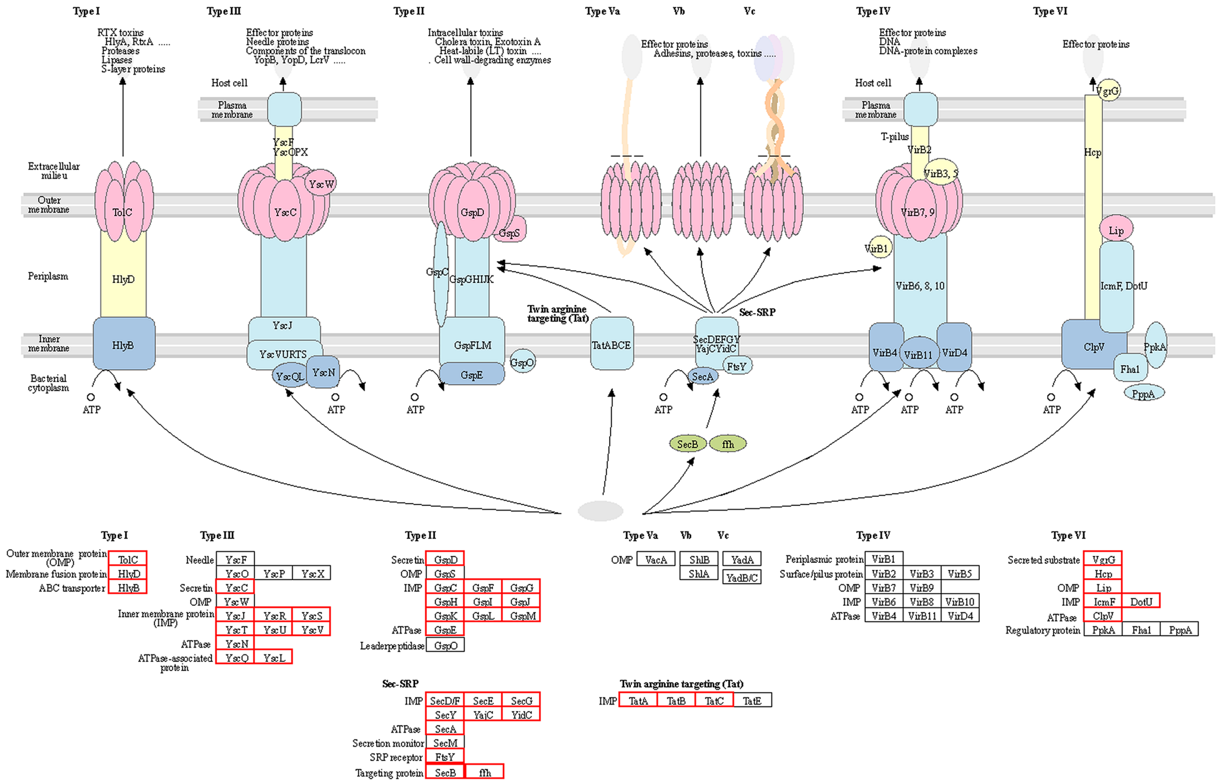


FIGURE 4
Comparison of genome circle analysis and plasmid comparative gene cluster analysis of gd-2. **(A)** Comparison of genome circle analysis of gd-2. (1) The large plasmid in the gd-2 genome is used as the reference genome, and the order from inside to outside is as follows: first circle: the GC content of the gd-2 plasmid; second circle: the GC skew of the gd-2 plasmid; third circle: the full-length sequence of the gd-2 plasmid; fourth circle: the full-length sequence of the B2 genome plasmid; fifth circle: the full-length sequence of the PhyI III-seqv23 genome chromosome III; sixth circle: the full-length sequence of the pMAFF241647 plasmid; seventh circle: the full-length sequence of the SL3822 genome plasmid; eighth circle: the full-length sequence of the FJAT15244.F50 genome Plas1 plasmid; ninth circle: the full-length sequence of the MP chromosome in the P824 genome; tenth circle: the full-length sequence of the plasmid in the OE1-1 genome; eleventh circle: the full-length sequence of the plasmid in the RS 476 genome; twelfth circle: the full-length sequence of the HA4-1MP plasmid; thirteenth circle: the mobile elements annotated on the gd-2 plasmid, including genomic islands, prophages, and CRISPR–Cas systems; fourteenth circle: the suspected drug resistance genes annotated on the gd-2 plasmid; and fifteenth circle: the suspected virulence genes annotated on the gd-2 plasmid. (2) Comparison of plasmid comparative gene cluster analysis of gd-2-PlasmidA. Sequence comparison analysis of GI12-GI14 of gd-2-PlasmidA with B2, 202, CQPS-1.



Type I

Outer membrane protein (OMP)	TolC
Membrane fusion protein	HlyD
ABC transporter	HlyB

Type III

Needle	YscF		
	YscO	YscP	YscX
Secretin	YscC		
OMP	YscW		
Inner membrane protein (IMP)	YscJ	YscR	YscS
	YscT	YscU	YscV
ATPase	YscN		
ATPase-associated protein	YscQ	YscL	

Type II

Secretin	GspD		
OMP	GspS		
IMP	GspC	GspF	GspG
	GspH	GspI	GspJ
	GspK	GspL	GspM
ATPase	GspE		
Leaderpeptidase	GspO		

Type Va Vb Vc

	Type Va	Vb	Vc
OMP	VacA	ShlB	YadA
		ShlA	YadB/C

Type IV

Periplasmic protein	VirB1		
Surface/pilus protein	VirB2	VirB3	VirB5
OMP	VirB7	VirB9	
IMP	VirB6	VirB8	VirB10
ATPase	VirB4	VirB11	VirD4

Type VI

Secreted substrate	VgrG		
	Hcp		
OMP	Lip		
IMP	IcmF	DotU	
ATPase	ClpV		
Regulatory protein	PpkA	Fha1	PppA

Sec-SRP

IMP	SecD/F	SecE	SecG
	SecY	YajC	YidC
ATPase	SecA		
Secretion monitor	SecM		
SRP receptor	FtsY		
Targeting protein	SecB	ffh	

Twin arginine targeting (Tat)

IMP	TatA	TatB	TatC	TatE

FIGURE 5
Statistical analysis of the structural composition of the secretion system of the gd-2 strain. The genes of each secretion system are represented by the gene names in the KEGG database, and the red boxes represent the corresponding genes in the strain.

with the number found in GMI1000 and CQPS-1. A comparative display analysis of the *hrp* gene cluster was performed for 4 samples, including GIM1000, CQPS-1, FQY-4, and gd-2 (Figure 6C).

4 Discussion

The *R. solanacearum* species complex (RSSC) has complex species types. According to the sequence data of the 16S-23S rRNA gene spacer region (ITS), *hrp* and *egl* genes, RSSC can be divided into four phylotypes (Phylotype I, II, III, and IV). The reported phylotype I *R. solanacearum* includes sequence variants 13, 14, 15, 17, 34, 44, 54, 55, etc. (Zheng et al., 2014; Liu et al., 2017). In Chian, sequevars present distribution difference at different tobacco planting zones, sequevars 15 has a high percent at in various planting zones, such as nanling hilly area (46.15%), wuyi hilly area (46.15%), huanghua plain area (100%) and yimeng hilly area (100%) (Liu et al., 2019). The genomic data of Ralstonia isolated from tobacco have been published, including Y45 (sequevar 17), FQY-4 (sequevar 17), CQPS-1 (sequevar 17), FJ1003 (sequevar 14) (Cao et al., 2013; Liu et al., 2017). Thus, we processed comprehensive genome sequence analysis of sequevar 15 to provide new evidence for ultimately analyzing the pathogenic specificity of *R. solanacearum* and the prevention and control of bacterial wilt. Through amplification and sequencing, gd-2 was identified as phylotype I sequevar 15. According to comparative genomic analysis, gd-2 maintains relative consistency with other strains in both the chromosomal and plasmid genomes. This study is the first reported whole-genome sequencing study of sequevar 15, providing new data in support of exploring the regulatory mechanism of virulence differentiation and host adaptation in *R. solanacearum*.

The composition and diversity of *R. solanacearum* groups caused by bacterial wilt are very complex. Nowdays, 55 sequence variants of *R. solanacearum* have been identified and the NCBI database has published the complete draft genome of 145 *R. solanacearum* (Ahn et al., 2011; Cai et al., 2015; Liu et al., 2017; Asolkar and Ramesh, 2018; Greenrod et al., 2023). The genome of *R. solanacearum* is approximately 5.8 Mb, dominated by two circular replicons, with the occasional presence of small plasmids, such as CMR15 containing a 35 kb small plasmid and PSI07 containing a 13 kb small plasmid (Liu et al., 2017). The genome size of gd-2 was 5.93 Mb, including the chromosomes (3.83 Mb) and the megaplasmid (2.10 Mb), which was larger than phylotype I strain FJ1003 (5.90 Mb), phylotype I sequevar 17 strain CQPS-1 (5.89 Mb) and phylotype I GMI1000 (5.8 Mb) (Salanoubat et al., 2002; Liu et al., 2017; Chen et al., 2022). Gene prediction results showed that 3,434 and 1,640 genes were identified in the chromosomes and plasmids of gd-2, which were similar with FJ1003 (3,446 chromosomes genes and 1,564 megaplasmid genes), CQPS-1 (3,573 chromosomes genes and 1,656 megaplasmid genes), phylotype I sequevar 14 M strain FJ1 (3,502 chromosomes genes and 1,596 megaplasmid genes) (Tan et al., 2022). The *hrp* gene cluster is an important component of the T3SS, which is necessary for the pathogenicity of *R. solanacearum* and can induce hypersensitivity reactions in non host plants (Lindgren, 1997). And 30 *hrp* genes identified both instrain gd-2 and strain CQPS-1. However, the

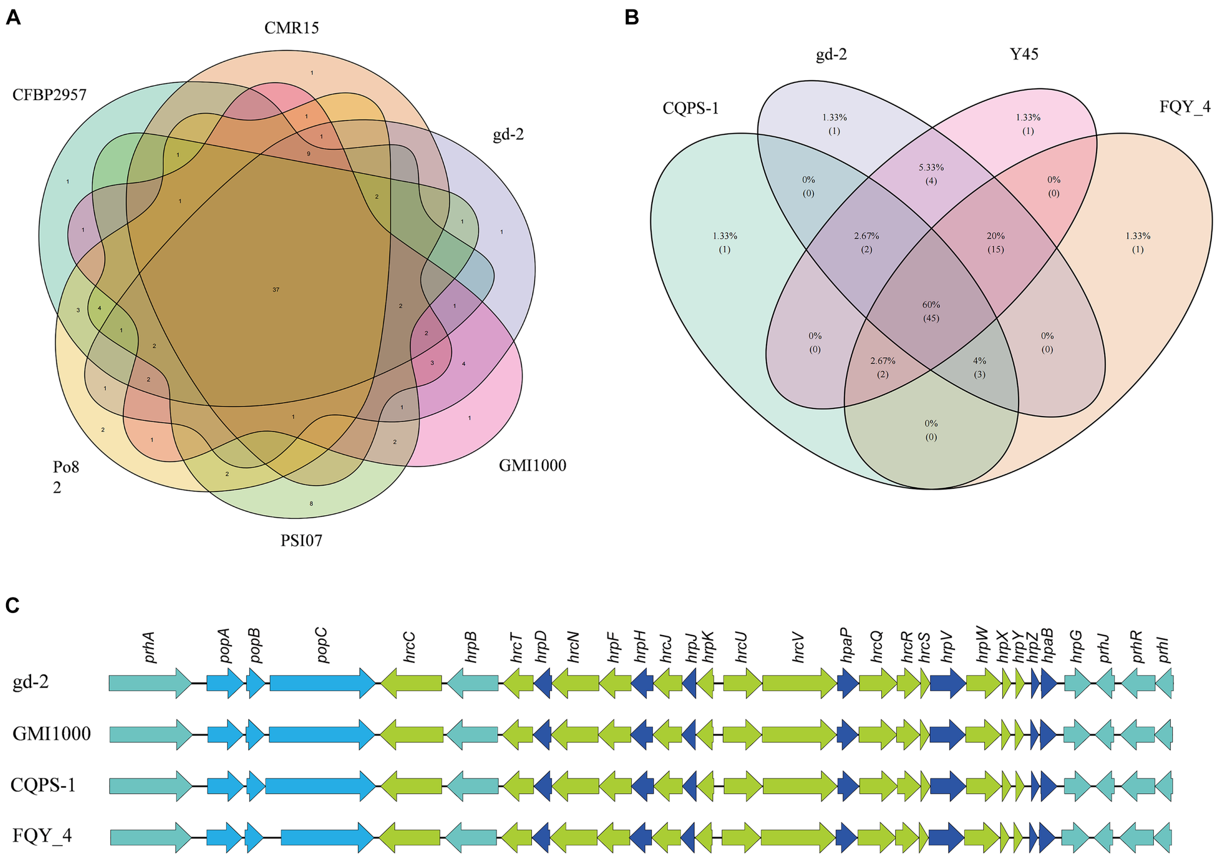


FIGURE 6
Venn diagram and gene comparison of *hrp* clusters among gd-2 and others *R. solanacearum* strain. **(A)** Venn diagram showing numbers of specific and shared gene families among gd-2 and five different phylotype species. **(B)** Venn diagram showing numbers of specific and shared gene families among gd-2 and three phylotype I species. **(C)** Genetic organization and gene comparison of *hrp* clusters between CQPS-1, Y45 and FQY_4.

numbers of GIs which is important forms of horizontal transfer elements of gd-2 was less than CQPS-1 (21), FJ1 (21), and FJ1003 (23), which may affect the adaptability of bacterial strain gd-2 to the environment. The predicted noncoding RNAs of gd-2 included 59 tRNAs, 12 rRNAs, four 5S rRNAs, four 16S rRNAs, and four 23S rRNAs. The number of tRNAs is similiar to CQPS-1 (58), biovar 4 Bs715 (59), FQY-4 (62), and FJ1 (59), significantly higher than FJ1003 (35), phylotype I YC45 (46), and Race 4 Biovar 4 SD54 (46) (Cao et al., 2013; Shan et al., 2013; She et al., 2015; Liu et al., 2017; Chen et al., 2022; Tan et al., 2022; Jeong et al., 2023). And more rRNA may improve protein synthesis ability and improve the adaptability of strains to the environment.

Pathogenic bacteria rely heavily on effector molecules secreted extracellularly or directly into host target cells to induce toxicity in the host or surrounding organisms. These different functional macromolecules are transported extracellularly through different secretion apparatuses (Li et al., 2012; Bai et al., 2018). Currently, seven types of secretion systems have been identified, which exhibit diversity not only in the effector molecules secreted but also in the composition of the apparatus. T1SS, T2SS, T3SS, T5SS, and T6SS are mainly found in gram-negative bacteria, while T7SS is mainly found in gram-positive bacteria. T4SS is found in both gram-positive and gram-negative bacteria (Parizad et al., 2016; Cordsmeier et al., 2022). T1SS and T5SS contain simple structure, consisting of only two or three proteins (Zhou et al., 2019). T2SS, T3SS, T4SS, and T6SS exist more complex than T1SS and T5SS and their apparatus can traverse the entire cell membrane (Korotkov et al., 2012). Study on T7SS is still in its infancy, and the specific apparatus and mechanism are still unclear (Bitter et al., 2009). T3SS, T4SS, and T6SS can directly inject effector molecules into eukaryotic cells, and they are mostly encoded by clusters of consecutive genes, especially in pathogenic bacteria, where these apparatus genes often exist as virulence islands (Liao et al., 2021). In this study, 66 secretion system structural genes were identified, including 5 in T1SS, 23 in T2SS, 10 in T3SS, 16 in T6SS, 11 in Sec-SRP, and 3 in Tat.

The T3SS effector proteins have an *hrp* II-box (TTCGN16-TTCG), which is activated by *HrpB* and *HrpG* transcription and powered by the ATPase complex. It enters plant cells through the cytoplasmic ring, basement, endomembrane exit, and transport pore (Mcnally et al., 2011). The effector proteins in *R. solanacearum* consist of 94 orthologous families, of which 71 are transferred or secreted through T3SS (Cunnac et al., 2004). A total of 72 T3SS proteins were identified in the strain gd-2 genome and 72 T3Es were identified, which is comparable to the number found in GMI1000 (74), FQY-4 (70), and Y45 (69). RipAZ2 is a unique T3E in gd-2 compared with other eight sequenced strain. However, there are significant differences between CQPS-1 (54) isolated from phylotype I of tobacco and CMR15 (61) isolated from phylotype III. The framework division of

R. solanacearum lineages result from the evolution and geographical origin of *R. solanacearum*, so it is speculated that the T3Es specific to each of the four *R. solanacearum* lineages may have been formed during the long-term evolution of the strains and their hosts, which also reflects the complexity of *R. solanacearum* species from another perspective.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession numbers PRJNA1071833 (BioProject) and SAMN39714303 (BioSample).

Author contributions

ZX: Methodology, Resources, Software, Validation, Visualization, Writing – original draft. GL: Formal analysis, Funding acquisition, Investigation, Writing – review & editing. AY: Formal analysis, Funding acquisition, Investigation, Writing – review & editing. ZL: Software, Visualization, Writing – review & editing. MR: Software, Visualization, Writing – review & editing. LC: Software, Visualization, Writing – review & editing. DL: Software, Visualization, Writing – review & editing. CJ: Conceptualization, Writing – review & editing. LW: Conceptualization, Writing – review & editing. SW: Conceptualization, Writing – review & editing. YC: Conceptualization, Writing – review & editing. WY: Resources, Writing – review & editing. RG: Resources, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by grants from the Natural Science Foundation of Tobacco Genome Project of State Tobacco Monopoly Administration: 110202201008 (JY-08); Agricultural Science and Technology Innovation Program: ASTIP-TRIC01.

Acknowledgments

The work was performed at the Key Laboratory for Tobacco Gene Resources, Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1335081/full#supplementary-material

References

Ahn, I., Lee, S., Gab, M., Sang-Ryeol, K., and Duk-ju, P. (2011). Priming by rhizobacterium protects tomato plants from biotrophic and necrotrophic pathogen infections through multiple defense mechanisms. *Mol. Cells* 32, 7–14. doi: 10.1007/s10059-011-2209-6

Alfano, J. R., and Collmer, A. J. (2004). Type III secretion system effector proteins: double agents in bacterial disease and plant defense. *Annu. Rev. Phytopathol.* 42, 385–414. doi: 10.1146/annurev.phyto.42.040103.110731

Asolkar, T., and Ramesh, R. J. (2018). Identification of virulence factors and type III effectors of phylotype I, Indian *Ralstonia solanacearum* strains Rs-09-161 and Rs-10-244. *J. Genet.* 97, 55–56. doi: 10.1007/s12041-018-0894-z

Bai, F., Li, Z., Umezawa, A., Terada, N., and Jin, S. J. (2018). Bacterial type III secretion system as a protein delivery tool for a broad range of biomedical applications. *Biotechnol. Adv.* 36, 482–493. doi: 10.1016/j.biotechadv.2018.01.016

Bitter, W., Houben, E., Bottai, D., Brodin, P., Brown, E. J., Cox, J. S., et al. (2009). Systematic genetic nomenclature for type VII secretion systems. *PLoS Pathog.* 5:e1000507. doi: 10.1371/journal.ppat.1000507

Buddenhagen, I., Sequeira, L., and Kelman, A. J. (1962). Designation of races in *Pseudomonas solanacearum*. *Phytopathology* 52:726.

Cai, L., Liu, Y., Meng, L., Luo, Z., and Shi, J. (2015). Bioinformatic analysis of prophage in genome of tobacco pathogen *Ralstonia solanacearum* FQY_4. *Acta Tabacaria Sin.* 21, 82–88. doi: 10.16472/j.chinatobacco.2013.505

Cao, Y., Tian, B., Liu, Y., Cai, L., Wang, H., Lu, N., et al. (2013). Genome sequencing of *Ralstonia solanacearum* FQY_4, isolated from a bacterial wilt nursery used for breeding crop resistance. *Genome Announc.* 1:e00125-13. doi: 10.1128/genomeA.00125-13

Castillo, J. A., and Greenberg, J. T. (2007). Evolutionary dynamics of *Ralstonia solanacearum*. *Appl. Environ. Microbiol.* 73, 1225–1238. doi: 10.1128/AEM.01253-06

Chen, K., Zhuang, Y., Wang, L., Li, H., Lei, T., Li, M., et al. (2022). Comprehensive genome sequence analysis of the devastating tobacco bacterial phytopathogen *Ralstonia solanacearum* strain FJ1003. *Front. Genet.* 13:966092. doi: 10.3389/fgene.2022.966092

Chen, T., Zhang, W. G., Zhu, H. J., Zeng, B. Y., Wang, R. E., Wang, X. Y., et al. (2020). Early detection of bacterial wilt in peanut plants through leaf-level hyperspectral and unmanned aerial vehicle data. *Comput. Electron. Agr.* 177:105708. doi: 10.1016/j.compag.2020.105708

Cheng, D., Zhou, D., Wang, Y., Wang, B., and Chen, H. (2021). *Ralstonia solanacearum* type III effector RipV2 encoding a novel E3 ubiquitin ligase (NEL) is required for full virulence by suppressing plant PAMP-triggered immunity. *Biochem. Biophys. Res. Commun.* 550, 120–126. doi: 10.1016/j.bbrc.2021.02.082

Choi, K., Son, G., Ahmad, S., Lee, S., and Lee, S. (2020). The plant pathology journal contribution of the *murI* gene encoding glutamate racemase in the motility and virulence of *Ralstonia solanacearum*. *Plant Pathol. J.* 36, 355–363. doi: 10.5423/PPJ.OA.03.2020.0049

Coll, N. S., and Valls, M. J. (2013). Current knowledge on the *Ralstonia solanacearum* type III secretion system. *Microb. Biotechnol.* 6, 614–620. doi: 10.1111/1751-7915.12056

Cordsmeier, A., Rinkel, S., Jeninga, M. D., Schulze-Luehrmann, J., Lke, M., Schmid, B., et al. (2022). The *Coxiella burnetii* T4SS effector protein AnkG hijacks the 7SK small nuclear ribonucleoprotein complex for reprogramming host cell transcription. *PLoS Pathog.* 18:e1010266. doi: 10.1371/journal.ppat.1010266

Cunnac, S., Occhialini, A., Barberis, P., Boucher, C., and Genin, S. (2004). Inventory and functional analysis of the large Hrp regulon in *Ralstonia solanacearum*: identification of novel effector proteins translocated to plant host cells through the type III secretion system. *Mol. Microbiol.* 53, 115–128. doi: 10.1111/j.1365-2958.2004.04118.x

Elphinstone, J., Allen, C., Prior, P., and Hayward, A. (2005). *The current bacterial wilt situation: a global overview. Bacterial wilt the disease & the Ralstonia Solanacearum species complex.*

Fegan, M., and Prior, P. (2005). "How complex is the *Ralstonia solanacearum* species complex?" in *Bacterial wilt disease and the Ralstonia solanacearum species complex*. eds. C. Allen, P. Prior and A. C. Hayward (St. Paul: American Phytopathological Society Press), 449–462.

Genin, S., and Denny, T. P. (2012). Pathogenomics of the *Ralstonia solanacearum* species complex. *Annu. Rev. Phytopathol.* 50, 67–89. doi: 10.1146/annurev-phyto-081211-173000

Greenrod, S. T., Stoycheva, M., and Elphinstone, J. (2023). Influence of insertion sequences on population structure of phytopathogenic bacteria in the *Ralstonia solanacearum* species complex. *Microbiology* 169:001364. doi: 10.1099/mic.0.001364

Guidot, A., Coupat, B., Fall, S., Prior, P., and Bertolla, F. (2009). Horizontal gene transfer between *Ralstonia solanacearum* strains detected by comparative genomic hybridization on microarrays. *ISME J.* 3, 549–562. doi: 10.1038/ismej.2009.14

Hayward, A. C. (2003). Biology and epidemiology of bacterial wilt caused by *Pseudomonas Solanacearum*. *Annu. Rev. Phytopathol.* 29, 65–87. doi: 10.1146/annurev.phyto.29.1.65

He, L. (1983). Characteristics of strains of *Pseudomonas solanacearum* from China. *Plant Dis.* 67, 1357–1361. doi: 10.1094/PD-67-1357

Huang, M., Tan, X., Song, B., Wang, Y., Cheng, D., Wang, B., et al. (2023). Comparative genomic analysis of *Ralstonia solanacearum* reveals candidate avirulence effectors in HA4-1 triggering wild potato immunity. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1075042

Ibtissem, G., Gilles, V., and Christine, P. J. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acid Res.* 35, 52–57. doi: 10.1093/nar/gkm360

Jeong, H., Ahn, H., Kim, S., Seol, Y., Yoon, H., Lee, J., et al. (2023). Complete genome sequence of *Ralstonia solanacearum* strain Bs715, a member of biovar 4 and a strong pathogen of bacterial wilt on *Solanum lycopersicum*. *Microbiol. Resour. Ann.* 12:e0088322. doi: 10.1128/mra.00883-22

Kang, Y. J., Li, S., and Zhao, L. (2008). Identification on biochemical variants of different *Ralstonia Solanacearum* on plant. *Biomol. Ther.* 11, 77–81. doi: 10.3390/biom11040560

Kang, Y., Liu, H., Genin, S., Schell, M. A., and Denny, T. P. (2002). *Ralstonia solanacearum* requires type 4 pili to adhere to multiple surfaces and for natural transformation and virulence. *Mol. Microbiol.* 46, 427–437. doi: 10.1046/j.1365-2958.2002.03187.x

Korotkov, K. V., Sandkvist, M., and Hol, W. G. (2012). The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* 10, 336–351. doi: 10.1038/nrmicro2762

Landry, D., Gonzalez-Fuente, M., Deslandes, L., and Peeters, N. (2020). The large, diverse and robust arsenal of *Ralstonia solanacearum* type III effectors and their in planta functions. *Mol. Plant Pathol.* 21, 1377–1388. doi: 10.1111/mpp.12977

Li, Z. G., He, F., Zhang, Z., and Peng, Y. L. (2012). Prediction of protein–protein interactions between *Ralstonia solanacearum* and *Arabidopsis thaliana*. *Amino Acids* 42, 2363–2371. doi: 10.1007/s00726-011-0978-z

Li, Y., Liu, H., Lin, W., Zhu, B., Huang, J., Xu, R., et al. (2015). Pathogenicity of *Ralstonia solanacearum* infecting tobacco in Enshi of Hubei province. *Chinese Tobacco Sci.* 36, 59–63. doi: 10.13496/j.issn.1007-5119.2015.05.011

Liao, W., Huang, H. H., Huang, Q. S., Fang, L. D., and Liu, Y. (2021). Distribution of type VI secretion system (T6SS) in clinical *Klebsiella pneumoniae* strains from a Chinese hospital and its potential relationship with virulence and drug resistance. *Microb. Pathog.* 162:105085. doi: 10.1016/j.micpath.2021.105085

Lindgren, P. B. (1997). The role of *hrp* genes during plant-bacterial interactions. *Annu. Rev. Phytopathol.* 35, 129–152. doi: 10.1146/annurev.phyto.35.1.129

Liu, Y., Tan, W., and Ding, W. (2019). Discussion on the subspecific diversity and unified naming of tobacco *Ralstonia solanacearum* in China. *Plant* 32, 15–17. doi: 10.13718/j.cnki.zwys.2019.06.002

Liu, Y., Tang, Y., Qin, X., Yang, L., Jiang, G., Li, S., et al. (2017). Genome sequencing of *Ralstonia solanacearum* CQPS-1, a phylotype I strain collected from a highland area with continuous cropping of tobacco. *Front. Microbiol.* 8:974. doi: 10.3389/fmicb.2017.00974

Liu, H., Zhang, S., Schell, M. A., and Denny, T. P. (2005). Pyramiding unmarked deletions in *Ralstonia solanacearum* shows that secreted proteins in addition to plant cell-wall-degrading enzymes contribute to virulence. *Mol. Plant Microbe Interact.* 18, 1296–1305. doi: 10.1094/MPMI-18-1296

Mao, L., Jiang, H., Wang, Q., Yan, D., and Cao, A. (2017). Efficacy of soil fumigation with dazomet for controlling ginger bacterial wilt (*Ralstonia solanacearum*) in China. *Crop Prot.* 100, 111–116. doi: 10.1016/j.cropro.2017.06.013

Mcnally, R., Sundin, G., Zhao, Y., Toth, I., and Hedley, P. (2011). Microarray characterization of the hrpl regulon of the fire blight pathogen Erwinia amylovora. *Acta horticulturae* 896, 263–270. doi: 10.17660/ActaHortic.2011.896.36

Mukaihara, T., Tamura, N., and Iwabuchi, M. J. (2010). Genome-wide identification of a large repertoire of *Ralstonia solanacearum* type III effector proteins by a new functional screen. *Mol. Plant Microbe Interact.* 23, 251–262. doi: 10.1094/MPMI-23-3-0251

Pan, X., Chen, J., Yang, A., Yuan, Q., Zhao, W., and Xu, T. (2021). Comparative transcriptome profiling reveals defense-related genes against *Ralstonia solanacearum* infection in tobacco. *Front. Plant Sci.* 12:767882. doi: 10.3389/fpls.2021.767882

Paret, M. L., Cabos, R., Kratky, B. A., and Alvarez, A. M. (2010). Effect of plant essential oils on *Ralstonia solanacearum* race 4 and bacterial wilt of edible ginger. *Plant Dis.* 94, 521–527. doi: 10.1094/PDIS-94-5-0521

Parizad, E. G., Parizad, E., Pakzad, I., and Valizadeh, A. (2016). A review of secretion systems in pathogenic and non-pathogenic bacteria. *Biosci. Biotechnol. Res. Asia* 13, 135–145. doi: 10.13005/bbra/2016

Peeters, N., Carrère, S., Anisimova, M., Plener, L., and Genin, S. J. (2013b). Repertoire, unified nomenclature and evolution of the type III effector gene set in the *Ralstonia solanacearum* species complex. *BMC Genomics* 6:859. doi: 10.1186/1471-2164-14-859

Peeters, N., Guidot, A., Vailleau, F., and Marc, V. (2013a). *Ralstonia solanacearum*, a widespread bacterial plant pathogen in the post-genomic era. *Mol. Plant Pathol.* 14, 651–662. doi: 10.1111/mpp.12038

Prokchorchik, M., Pandey, A., Moon, H., Kim, W., Jeon, H., Jung, G., et al. (2020). Host adaptation and microbial competition drive *Ralstonia solanacearum* phylotype I evolution in the Republic of Korea. *Microb. Genom.* 6:mgen000461. doi: 10.1099/mgen.0.000461

Qian, Y. L., Wang, X. S., Wang, D. Z., Zhang, L. N., and Yao, D. N. (2012). The detection of QTLs controlling bacterial wilt resistance in tobacco (*N. tabacum* L.). *Euphytica* 192, 259–266. doi: 10.1007/s10681-012-0846-2

Ran, G., Zhu, C., Tian, Y., Guo, P., Huang, Y., and Xiao, J. (2014). Solid formulation of hrp-mutant of *Ralstonia solanacearum* as biocontrol agent of bacterial wilt. *Chinese Journal of Biological Control* 30, 385–392. doi: 10.1097/MOP.0b013e3283423f35

Sabbagh, C. R., Carrère, S., Lonjon, F., Vailleau, F., and Peeters, N. (2019). Pangenomic type III effector database of the plant pathogenic *Ralstonia* spp. *Peer J.* 7:e7346. doi: 10.7717/peerj.7346

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., et al. (2002). Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* 415, 497–502. doi: 10.1038/415497a

Schachterle, J. K., and Huang, Q. J. (2021). Implication of the type III effector RipS1 in the cool-virulence of *Ralstonia solanacearum* strain UW551. *Front. Plant Sci.* 12:1430. doi: 10.3389/fpls.2021.705717

Shan, W., Yang, X., Ma, W., Yang, Y., Guo, X., Guo, J., et al. (2013). Draft genome sequence of *Ralstonia solanacearum* race 4 biovar 4 strain SD54. *Genome Announc.* 1:e00890-13. doi: 10.1128/genomeA.00890-13

Sharma, T. P. (2021). Genome resource: *Ralstonia solanacearum* phylotype II sequevar 1 (race 3 biovar 2) strain UW848 from the 2020 U.S. geranium introduction. *Plant Dis.* 105, 207–208. doi: 10.1094/PDIS-06-20-1269-A

Sharma, K., Kreuze, J., Abdurahman, A., Parker, M. L., and Rukundo, P. J. (2020). Molecular diversity and pathogenicity of *Ralstonia solanacearum* species complex associated with bacterial wilt of potato in Rwanda. *Plant Dis.* 105, 770–779. doi: 10.1094/PDIS-04-20-0851-RE

She, X., Tang, Y., He, Z., and Lan, G. (2015). Genome sequencing of *Ralstonia solanacearum* race 4, biovar 4, and phylotype I, strain YC45, isolated from *Rhizoma kaempferiae* in southern China. *Genome Announc.* 3:e01110-15. doi: 10.1128/genomeA.01110-15

Shrivastava, S., Reddy, C., and Mande, S. (2010). INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J. Biosci.* 35, 351–364. doi: 10.1007/s12038-010-0040-4

Stritzler, M., Soto, G., and Ayub, N. (2018). Plant growth-promoting genes can switch to be virulence factors via horizontal gene transfer. *Microb. Ecol.* 76, 579–583. doi: 10.1007/s00248-018-1163-7

Tan, X., Dai, X., Chen, T., Wu, Y., Yang, D., Zheng, Y., et al. (2022). Complete genome sequence analysis of ralstonia solanacearum strain PeaFJ1 provides insights into its strong virulence in peanut plants. *Front. Microbiol.* 13:830900. doi: 10.3389/fmicb.2022.830900

Tan, X., Qiu, H., Li, F., Cheng, D., and Xie, C. J. (2019). Complete genome sequence of sequevar 14M *Ralstonia solanacearum* strain HA4-1 reveals novel type III effectors acquired through horizontal gene transfer. *Front. Microbiol.* 10:1893. doi: 10.3389/fmicb.2019.01893

Tasset, C., Bernoux, M., Jauneau, A., Pouzet, C., Brière, C., Kieffer-Jacquinod, S., et al. (2010). Autoacetylation of the *Ralstonia solanacearum* effector PopP2 targets a lysine residue essential for RRS1-r-mediated immunity in Arabidopsis. *PLoS Pathog.* 6:e1001202. doi: 10.1371/journal.ppat.1001202

Tsai, A. Y., English, B. C., Wiley, R. M., and Sons, L. (2019). Hostile takeover: hijacking of endoplasmic reticulum function by T4SS and T3SS effectors creates a niche for intracellular pathogens. *Microbiol. Spectr.* 7:10. doi: 10.1128/microbiolspec.PSIB-0027-2019

Tsujimoto, S., Nakaho, K., Adachi, M., Ohnishi, K., Kiba, A., and Hikichi, Y. J. (2008). Contribution of the type II secretion system in systemic infectivity of *Ralstonia solanacearum* through xylem vessels. *J. Gen. Plant Pathol.* 74, 71–75. doi: 10.1007/s10327-007-0061-5

Valls, M., Genin, S., and Boucher, C. J. (2006). Integrated regulation of the type III secretion system and other virulence determinants in *Ralstonia solanacearum*. *PLoS Pathog.* 2:e82. doi: 10.1371/journal.ppat.0020082

Yuliar, N., Yanetri, A., and Toyota, K. (2015). Recent trends in control methods for bacterial wilt diseases caused by *Ralstonia solanacearum*. *Microbes Environ.* 30, 1–11. doi: 10.1264/jsme2.ME14144

Zheng, X., Zhu, Y., Liu, B., Zhou, Y., Wang, J., Zhang, H., et al. (2014). Relationship between *Ralstonia solanacearum* diversity and severity of bacterial wilt disease in tomato fields in China. *Can. J. Microbiol.* 162, 607–616. doi: 10.1139/cjm-2018-0637

Zhou, D., Wang, S., Wu, X., Yi, Z., Xin, S., Zhang, Y., et al. (2019). Distribution and epidemiological analysis of type V secretion system (T5SS) in avian pathogenic *Escherichia coli*. *Microbiol. China* 46, 3076–3083. doi: 10.13344/j.microbiol.china.190446

TYPE Original Research
PUBLISHED 28 May 2024
DOI 10.3389/fgene.2024.1368843

OPEN ACCESS

EDITED BY
Tulio de Lima Campos,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Tiago Gräf,
Instituto Carlos Chagas (ICC), Brazil
Camila Malta Romano,
University of São Paulo, Brazil

*CORRESPONDENCE
Lee Ching Ng,
✉ ng_lee_ching@nea.gov.sg
Hapuarachchige Chanditha Hapuarachchi,
✉ Chanditha_hapuarachchi@nea.gov.sg

†These authors have contributedequally to this work

RECEIVED 11 January 2024
ACCEPTED 16 April 2024
PUBLISHED 28 May 2024

CITATION
Chem YK, Yenamandra SP, Chong CK, Mudin RN, Wan MK, Tajudin N, Abu Bakar RS, Yamin MA, Yahya R, Chang C-C, Koo C, Ng LC and Hapuarachchi HC (2024), Molecular epidemiology of dengue in Malaysia: 2015–2021.
Front. Genet. 15:1368843.
doi: 10.3389/fgene.2024.1368843

Molecular epidemiology of dengue in Malaysia: 2015–2021

Yu Kie Chem[1†], Surya Pavan Yenamandra[2†], Chee Keong Chong[3], Rose Nani Mudin[3], Ming Keong Wan[3], Norazimah Tajudin[1], Rehan Shuhada Abu Bakar[1], Mohd Asri Yamin[1], Rokiah Yahya[1], Chia-Chen Chang[2], Carmen Koo[2], Lee Ching Ng[2,4,5]* and Hapuarachchige Chanditha Hapuarachchi*

[1]National Public Health Laboratory, Ministry of Health, Sungai Buloh, Malaysia, [2]Environmental Health Institute, National Environment Agency, Singapore, Singapore, [3]Disease Control Division, Ministry of Health, Putrajaya, Malaysia, [4]School of Biological Sciences, Nangyang Technological University, Singapore, Singapore, [5]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Dengue has been one of the major public health problems in Malaysia for decades. Over 600,000 dengue cases and 1,200 associated fatalities have been reported in Malaysia from 2015 to 2021, which was 100% increase from the cumulative total of dengue cases reported during the preceding 07-year period from 2008 to 2014. However, studies that describe the molecular epidemiology of dengue in Malaysia in recent years are limited. In the present study, we describe the genetic composition and dispersal patterns of Dengue virus (DENV) by using 4,004 complete envelope gene sequences of all four serotypes (DENV-1 = 1,567, DENV-2 = 1,417, DENV-3 = 762 and DENV-4 = 258) collected across Malaysia from 2015 to 2021. The findings revealed that DENV populations in Malaysia were highly diverse, and the overall heterogeneity was maintained through repetitive turnover of genotypes. Phylogeography analyses suggested that DENV dispersal occurred through an extensive network, mainly among countries in South and East Asia and Malaysian states, as well as among different states, especially within Peninsular Malaysia. The results further suggested Selangor and Johor as major hubs of DENV emergence and spread in Malaysia.

KEYWORDS

dengue, molecular epidemiology, Malaysia, genetic diversity, phylogeography. severity

1 Introduction

Dengue has long been one of the major public health problems in Malaysia. Dengue fever and dengue hemorrhagic fever were first documented in Penang, a northern state of Malaysia, in 1902 and 1962 respectively (Skae, 1902; Rudnick et al., 1965). However, the disease became noticeable since early 1970s (Lam, 1993) after the first major epidemic of severe dengue due to Dengue virus type 3 (DENV-3) in 1973 (Wallace et al., 1980; Mudin, 2015). Subsequent epidemics occurred in a cyclical pattern of approximately 5–8 years [10], mainly affecting young adults living in urbanized areas, such as Selangor, Johor and Kuala Lumpur (Mohd-Zaki et al., 2014). Dengue incidence continued to increase dramatically after 2001, largely due to the factors associated with rapid urbanization (Abubakar and Shafee, 2002). Based on 2010 estimates, Malaysia spent over US$175.7 million annually on dengue prevention and control (Packierisamy et al., 2015).

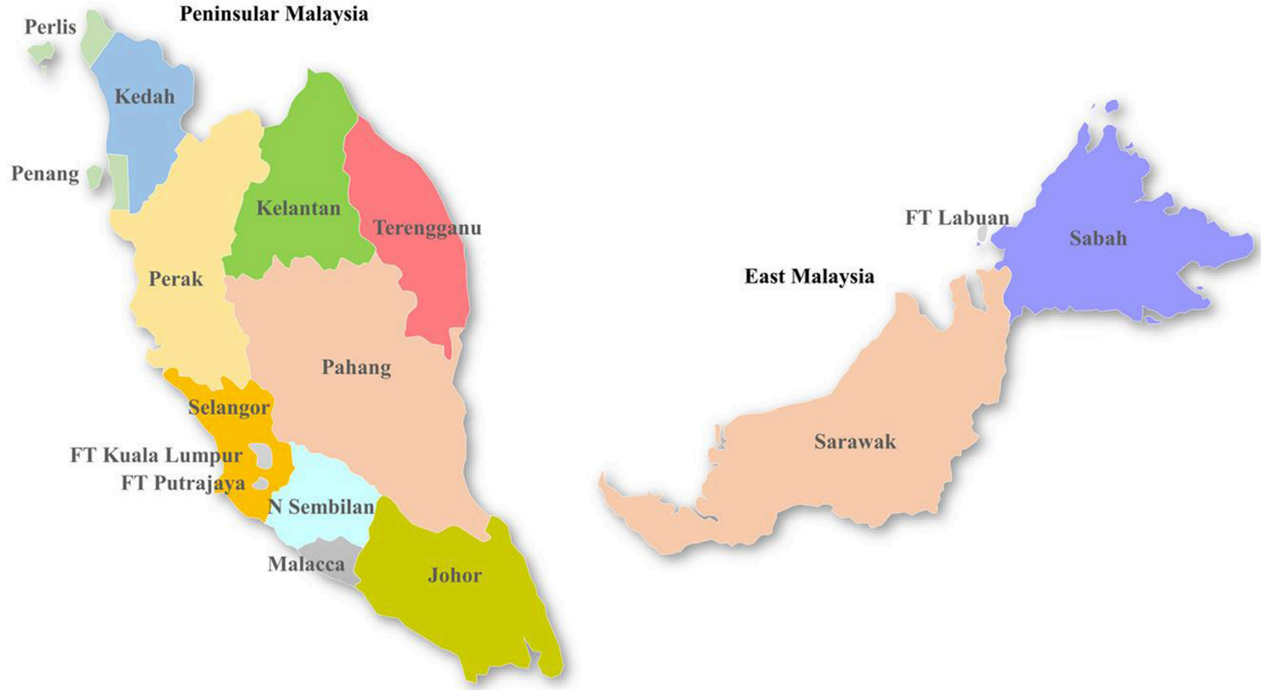


FIGURE 1
Map of Malaysia, illustrating different states and Federal Territories (FT).

Dengue is hyperendemic in Malaysia, with the co-circulation of all four DENV serotypes (AbuBakar et al., 2022). DENV-1 was first reported in Kuala Lumpur in 1954 (Smith, 1956), followed by other serotypes. Major epidemics have mainly been due to DENV-1, -2 and -3, with DENV-4 at a relatively low intensity since 1969 (Abubakar and Shafee, 2002). None of the serotypes was clearly dominant until mid-1980s but DENV-3 became dominant in 1986, followed by DENV-1, and -2 in subsequent years (Abubakar and Shafee, 2002; Mohd-Zaki et al., 2014; Ng et al., 2015). Even though DENV-4 has historically been the least common serotype, it showed a rapid upsurge during 2021–2022 (Suppiah et al., 2023). Based on the previous studies, genotypes I and II of DENV-1 have been associated with major outbreaks in Klang valley during 1987–2004 (Teoh et al., 2013), whereas DENV-2 cosmopolitan genotype caused the epidemics during 1989–2000 (Chee and AbuBaker, 2003) and in 2013 (Ng et al., 2015).

Despite the historical presence of DENV, very few studies have narrated the molecular epidemiology of dengue in Malaysia, especially in the last decade during which cases escalated further. Understanding the temporal dynamics of DENV can provide important insights into dengue epidemiology and outbreak risk assessment (Lee et al., 2010) that are vital elements of outbreak control. In this descriptive study, we therefore analyzed the genetic, evolutionary and dispersal characteristics of DENV serotypes in Malaysia from 2015 to 2021. The findings revealed a highly heterogeneous DENV population that dispersed through an extensive network, involving important hubs of virus emergence, and spread in Malaysia.

2 Materials and methods

2.1 Sample collection

Malaysia is administratively divided into 13 states and three Federal Territories, and physically separated into two regions by the South China Sea: Peninsular Malaysia and East Malaysia (Figure 1). Approximately 79% of Malaysia's ~30 million population lives in Peninsular Malaysia, mainly concentrated in urban areas.

All sera used to generate genome sequences in the present study were obtained through a national dengue virus surveillance program that screened in- and outpatients who sought treatment at 52 sentinel hospitals and clinics throughout the country. The clinical severity of these infections was classified into dengue fever with and without warning signs and severe dengue, according to the 2009 World Health Organization (WHO) guidelines (WHO, 2009). Severe dengue manifestations included severe plasma leakage, severe hemorrhage, and severe organ dysfunction. Serum samples obtained from suspected dengue patients were tested for dengue by using either NS1 antigen rapid test or ELISA assay. In each sentinel site, five DENV-NS1 antigen positive serum samples were randomly selected every week and were sent to the National Public Health Laboratory (NPHL) and four other regional Public Health Laboratories (PHL) for reverse transcription quantitative polymerase chain reaction (RT-qPCR) to determine DENV serotypes.

Genome sequencing was also done as part of the same dengue surveillance program. These genome sequences were shared in the online data sharing platform UNITEDengue (United in Tackling Epidemic Dengue; www.unitedengue.org), which was launched in August 2012 to support cross-border surveillance and capacity building for dengue. The analyses of shared data allow visualization of dengue incidence trend and changes in dominant serotypes and genotypes between neighboring countries. Disease Control Division of the Ministry of Health, Malaysia is a founding partner of UNITEDengue.

2.2 Serotyping of dengue virus

Dengue virus RNA was extracted from DENV NS1 positive sera using the NucleoSpin® RNA Virus Kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's recommendations. Dengue virus serotypes were determined by using the abTES™ Kit (AITBiotech, Singapore) as per the manufacturer's recommended protocol and guidelines.

2.3 Sequencing of envelope gene of dengue virus serotypes, clade classification and relatedness analysis to global sequences

Complementary DNA (cDNA) was synthesized by using extracted viral RNA and random hexamer primers, according to the protocol recommended for SuperScript III first-strand synthesis system kit (Life Technologies, Carlsbad, United States). The complete *E* gene of each serotype was amplified by using serotype-specific primers as per the protocol described elsewhere (Lee et al., 2012; Sy et al., 2023). PCR amplicons (DENV-1 = 1,780 bp; DENV-2 = 1,770 bp; DENV-3 = 1,711 bp; DENV-4 = 1,907 bp) were visualized in 1% agarose gels and were purified by using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany) prior to sequencing at a commercial facility according to the BigDye Terminator Cycle Sequencing kit protocol (Applied Biosystems, United States). Consensus sequences for each sample were derived by assembling the overlapping raw nucleotide data in the Lasergene version 15.0 (DNASTAR Inc. Madison, WI, United States). The consensus sequences were aligned by using ClustalW implemented in BioEdit 7.2.5 software suite (Hall, 1999). The final dataset included 5,471 *E* gene sequences (4,004 complete and 1,468 partial sequences) belonging to four DENV serotypes (DENV-1 = 1,567, DENV-2 = 1,417, DENV-3 = 762 and DENV-4 = 258) reported in all states, except Labuan. This dataset was analyzed together with 17,732 *E* gene sequences (DENV-1 = 8,323, DENV-2 = 5,207, DENV-3 = 2,563 and DENV-4 = 1,639) retrieved from the GenBank database to identify distinct clades among study sequences and to determine their closest kins from other countries and thereby possible sources of their introductions by using the maximum likelihood (ML) method implemented in MEGA7 software suite (Kumar et al., 2016). Tamura-Nei model with gamma correction (TN93 + G_4+I) was selected as the best-fit model by jModel Test for DENV-1, -2 and -3 and General Time Reversal (GTR + G_4+I) model for DENV-4 (Darriba et al., 2012).

2.4 Bayesian phylogenetic and phylogeography analyses of envelope gene sequences

Bayesian time scaled phylogeny and phylogeography analyses were conducted separately for each serotype by using complete *E* gene sequences in Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software package v1.10.4 (Suchard et al., 2018). The phylogenetic analyses were carried out to estimate the most probable origin, nucleotide substitution rate and the time to most recent common ancestor (tMRCA) of different DENV lineages circulated in Malaysia during the study period. The dataset, therefore included Malaysian sequences representative of all clades and those outside of any clade (non-clades). The global dataset included sequences that were interspersed within and basal to Malaysian clades and those closely related to non-clades. The global sequences were selected based on their genetic relatedness to study sequences determined by using the ML tree constructed with all study and global sequences. The final dataset used for phylogenetic analyses included 1,417 Malaysian sequences (DENV-1 = 379; DENV = 411; DENV-3 = 388; DENV-4 = 239), and 459 global sequences (DENV-1 = 147, DENV-2 = 120, DENV-3 = 119 and DENV-4 = 73).

The discrete phylogeography approach was conducted by using a symmetrical Bayesian Stochastic Search Variable Selection (BSSVS) component (Lemey et al., 2009) to understand dispersal pattern of DENV in Malaysia driven by inter-state migration and external introductions from other countries. Unlike in phylogenetic analyses, phylogeography analyses included sequences of only the most common clades that represented >90% of all sequences of each serotype (DENV-1: clades A, C and D; DENV-2: clade Ib; DENV-3: clades A, C and D; DENV-4: clades A and B) and sequences from other countries that were interspersed within each selected clade. The selected clades included 95% (n = 3,810) of 4,004 complete *E* gene sequences generated, and were therefore the main contributors of overall DENV transmission in Malaysia during the study period. The final dataset used for phylogeography analyses included 1,232 study sequences (DENV-1 = 312; DENV = 325; DENV-3 = 360; DENV-4 = 235) and 369 sequences from other countries (DENV-1 = 164; DENV-2 = 133; DENV-3 = 46; DENV-4 = 26). Malaysian sequences were categorized into 14 discrete clusters based on case locations in respective administrative states. The size of each complete dataset was trimmed by removing sequence redundancy based on each state/country and reported year to optimize computational load.

The temporal signal of final datasets for each serotype was tested by using TempEst version 1.5.3 (Rambaut et al., 2016). Tamura-Nei model with gamma correction (TN93 + G_4+I) was used as the substitution model for DENV-1, -2 and -3 and GTR + G_4+I for DENV-4 (Darriba et al., 2012). In order not to assume any particular demographic scenario as *a priori*, a relaxed uncorrelated lognormal clock and the Bayesian skyline plot (10-steps) coalescent model (Lemey et al., 2009) were used. The MCMC chains were run for 100 million generations sampling every 10,000 states. The output log files were visualized in Tracer v.1.5 (Rambaut et al., 2018), and Effective Sampling Size (ESS) of >200 was considered as a sufficient level of convergence of parameters. The maximum clade credibility tree (MCC) was constructed after removing the first 10% of all trees (burn-in) by using TreeAnnotator v.1.7.4. The MCC tree was visualized in FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). Phylogeographic reconstructions and visualization of diffusion links were conducted in SpreaD3 0.9.6 (Bielejec et al., 2016). In SpreaD3 analyses, all sequences from a particular state (discrete trait) were assigned a common latitude and longitude value that represents the center of respective state. Significant migration links between discrete traits were determined by using Bayes factor (BF) values. BF > 3 was considered well-supported, with sub-classifications of substantial (BF > 3), strong (BF > 10), very strong (BF > 30) and decisive (BF > 100) support (Faria et al., 2013). The root state posterior probability values for each discrete trait were extracted from the annotated MCC trees obtained from phylogeography analyses to determine the most probable location of origin of different virus lineages.

2.5 Statistical analysis

The correlation between yearly dengue cases and the proportion of serotypes in each year from 2015 to 2021 was determined by performing univariate linear regression between transformed dengue cases (centered to mean) and the proportion of serotype each year. Moreover, the association between the genotype and severity of infection was determined by using a logistic regression model with quasibinomial error structure to account for overdispersion. The binary response variable was severe (=1) or mild (=0) infection. The predictor included genotype (DENV-1 genotype I as reference due to the most common genotype) while controlling for age, and sex (binary) of a patient. All analyses were done using R version 4.3.1 (RCoreTeam, 2018).

3 Results

3.1 Dengue cases were mainly concentrated in Klang valley and Johor state during the study period from 2015 to 2021

Malaysia reported 120,836 dengue cases (389 cases/100,000 population), inclusive of 336 mortalities in 2015. The case burden trended downwards from 2016 to 2018 but escalated in 2019 to record the highest case burden (130,101 cases; 397 cases/100,000 population) reported since 2004 (Figure 2A). Even though cases trended downwards in 2020 and 2021, presumably due to COVID-19 lockdown effects (Ahmad Zaki and Xin, 2023; Iderus et al., 2023), the average annual case burden almost doubled during the period from 2015 to 2021 (90,125 cases; average incidence of 280 cases/100,000 population) as compared to the preceding 7-year period from 2008 to 2014 (~47,000 cases; average incidence of 161 cases/100,000 population). The highest case burden during the study period was in Selangor state (53.2%), followed by Johor state (10.5%) and Federal Territory Kuala Lumpur (9.8%) (Figure 2B). Monthly cases in Selangor state closely and consistently followed the nationwide case pattern. This is not surprising because case burden has historically been high in these regions (Hassan et al., 2012) where highly urbanized and populated cities such as Klang, Petaling Jaya, Shah Alam, Kajang and Subang as well as Federal Territories Kuala Lumpur and Putrajaya are located (Figure 1). The case fatality rate (CFR) averaged at 0.2% during the 2015–2021 period and was strongly and positively correlated with dengue case numbers (Pearson's r = 0.82, $p = 0.02$). Of 3,331 severe dengue infections (inclusive of 1,244 fatalities) reported in different states from 2015 to 2021, the highest proportions were recorded in Selangor (47.6%), Federal Territory Kuala Lumpur and Putrajaya (18.7%) and Johor (6.6%) states. The highest proportions among all dengue-related fatalities across Malaysia were also reported in the same states (Selangor-32.4%; Johor-15.6% and Federal Territory Kuala Lumpur and Putrajaya-8.3%) during the study period. Interestingly, the proportion of severe infections continued to drop from 2015 to 2021, despite having the highest case burden in 2019 (Figures 2A,B).

3.2 Dengue was hyperendemic, but the distribution of common serotypes varied spatially and temporally during the study period

Based on the analysis of 42,763 clinical samples, all DENV serotypes circulated in Malaysia during the study period (Figure 3A). DENV-2 (n = 15,328; 35.9%) and DENV-1 (n = 14,302; 33.4%) were the most common serotypes, followed by DENV-3 (n = 11,678; 27.3%) and DENV-4 (n = 1,455; 3.4%). However, the proportion of each serotype fluctuated over time and one serotype was dominant periodically - DENV-1 in 2015 (53.1%) and 2016 (40.3%), DENV-3 in 2017 (41.1%) and 2020 (37%), DENV-2 in 2018 (47%) and 2019 (42.4%) and DENV-4 in 2021 (32%). None of the serotypes exceeded 50% of the total serotyped cases, except in 2015, indicating no clear dominance of a single serotype across Malaysia from 2016 to 2021. There was also no statistically significant correlation between the total number of cases reported and the proportion of serotypes across whole Malaysia in each year (DENV-1: Est = 2.68, SE = 3.16, p-value = 0.44; DENV-2: Est = 5.85, SE = 3.41, p-value = 0.15; DENV-3: Est = −1.06, SE = 4.33, p-value = 0.82; DENV-4: Est = −6.67, SE = 2.14, p-value = 0.03). It is important to note that the negative association between DENV-4 and yearly dengue cases is mostly driven by 2021 (after excluding 2021, Est = −1.61, SE = 8.70, p-value = 0.86). This lack of association between proportion of serotype and dengue cases was because the distribution of serotypes and their proportions differed among different states in each year (Figure 3B, Supplementary Figure S5). For example, DENV-1 was the most common serotype in all states in 2015, except Sabah and Sarawak, where DENV-2 was more common. Similarly, DENV-2 was the dominant serotype in many states in 2018, but DENV-1 and DENV-3 were more common in Penang and Sabah respectively. In general, the dominant serotype pattern was different from that of peninsular Malaysia in states such as Sabah, Sarawak and Penang that are physically distant from peninsular Malaysia (Figure 1). On the other hand, a single serotype often dominated in different states in each year, but such dominance did not show a statistically significant correlation with the total number of cases reported in each state in respective years (Supplementary Figure S5).

3.3 Heterogeneity of DENV populations fluctuated in time and space and different lineages dispersed through an extensive network across Malaysia during 2015–2021

The mean nucleotide substitution rate (subs/site/year) of each serotype did not differ substantially [DENV-1 = 7.4×10^{-4} (95% HPD = 6.6×10^{-4}-8.2×10^{-4}), DENV-2 = 9.0×10^{-4} (95% HPD = 8.0×10^{-4}-9.9×10^{-4}), DENV-3 = 8.8×10^{-4} (95% HPD = 7.7×10^{-4}-9.9×10^{-4}) and DENV-4 = 8×10^{-4} (95% HPD = 6.3×10^{-4}-9.8×10^{-4})]. The inferred median root ages of each serotype ranged from ~107.18 years (DENV-3) to ~256.81 years (DENV-2), which are

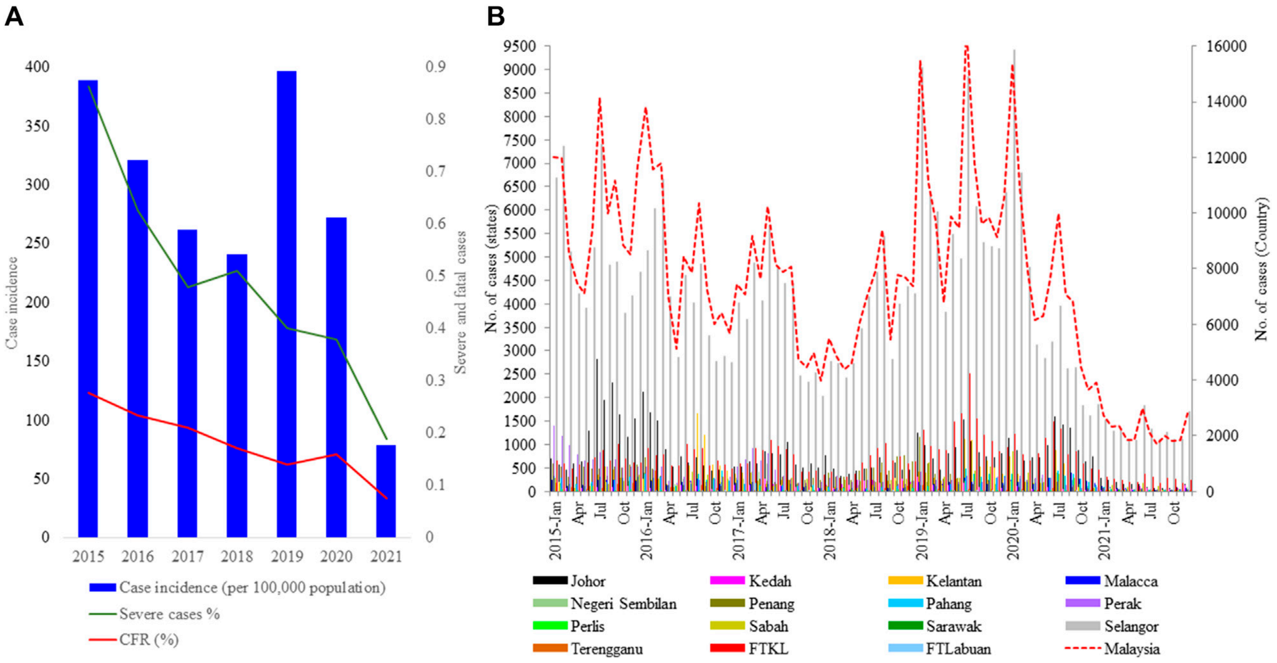


FIGURE 2
Dengue case burden in Malaysia from 2015 to 2021. **(A)**. Case incidence and percentage of severe infections and fatalities **(B)**. Distribution of dengue cases in different states. FTKL = Federal Territory Kuala Lumpur; FTLabuan = Federal Territory Labuan; FT Putrajaya = Federal Territory Putrajaya; N Sembilan = Negeri Sembilan. Population data for the calculation of annual incidence was obtained from https://data.worldbank.org.

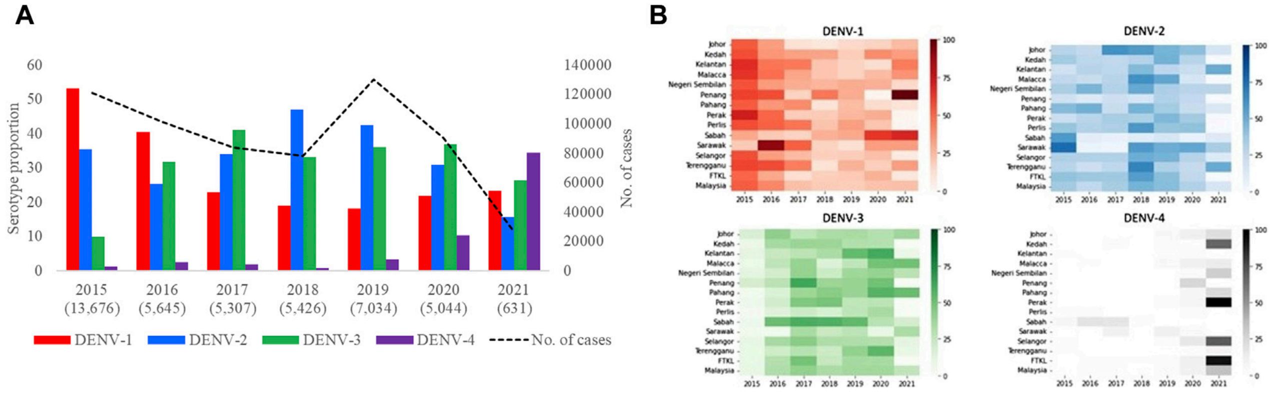


FIGURE 3
Distribution of DENV serotypes in Malaysia from 2015 to 2021. **(A)**. Temporal fluctuation of DENV serotypes in Malaysia (overall) and **(B)**. different states of Malaysia in each year. The fluctuations of all serotypes in each state are given separately in Supplementary Figure S5. Numbers within brackets in Figure 3A are the total number of samples subjected to serotyping. Substantially lower number of samples were serotyped in 2021 due to the low case burden. The legend in Figure 3B represents proportions of each serotype in different shades of respective colours. There was no serotype data from Federal Territory Labuan. FTKL = Federal Territory of Kuala Lumpur.

comparable to previous estimates (Wolf et al., 2023). Among 5,471 DENV *E* gene sequences analysed, a single genotype was largely dominant in each serotype (DENV-1 genotype I, DENV-2 cosmopolitan genotype, DENV-3 genotype III and DENV-4 genotype II) during the study period (Figure 4).

However, there were 18 distinct monophyletic clades belonging to different genotypes of each serotype, with strong posterior probability support (>0.9) (Table 1; Supplementary Figures S1-S4). Ancestral analyses (tMRCA) suggested long-term survival of all those clades (Table 1). All clades had basal sequences reported from South and East Asian countries as well as Australia (Supplementary Figures S1-S4). Each clade included sequences from many regional countries, suggesting the widespread presence of respective clades in South and East Asia and indicating the potential sources of their introduction into Malaysia (Supplementary Table S1). Even though 98.8% (1945/1969) of DENV-1 sequences belonged to genotype I, it included four distinct clades that were likely to have co-circulated,

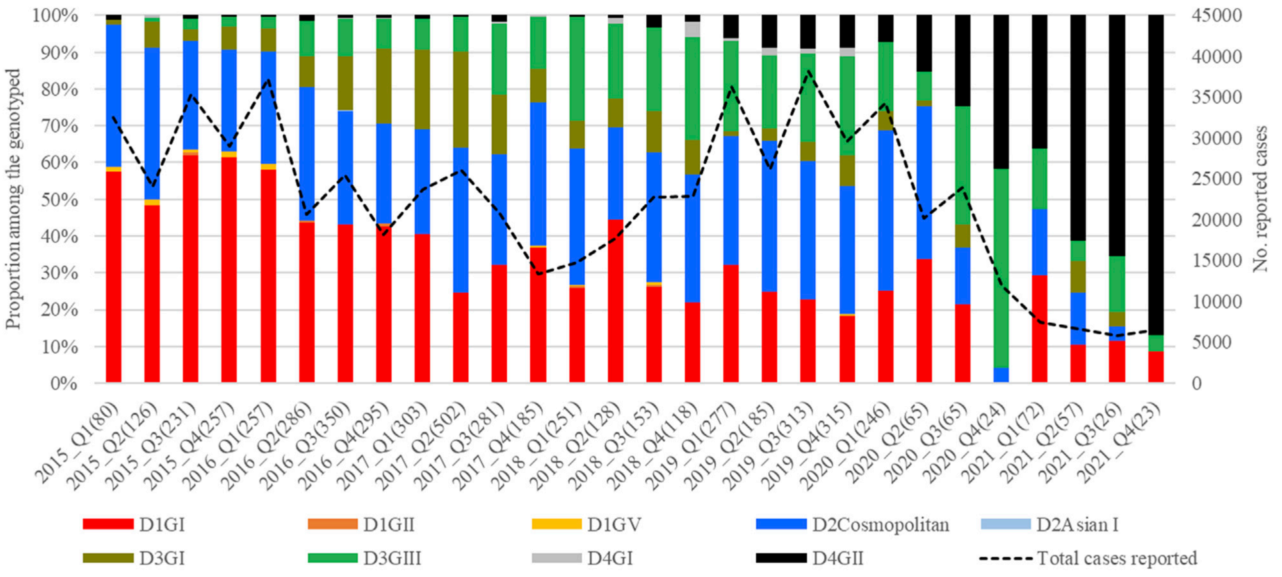


FIGURE 4
Temporal pattern of DENV genotype composition in Malaysia: 2015–2021. Each colour represents a genetically distinct genotype belonging to different serotypes. D1-D4 = DENV-1 to DENV-4; GI-GIII = Genotype I to III, Q1-Q4 = Quarter one to four.

demonstrating a relatively higher heterogeneity in DENV-1 than in other serotypes. Any one of those clades dominated at a time, suggesting a consistent lineage turnover within DENV-1 (Supplementary Figure S6). Notably, DENV-1 genotype I clade D that contributed to high case burden in 2015 was replaced by clade A in Q1-2016, before a decline in the overall proportion of DENV-1 by mid-2016. This replacement coincided with an overall reduction in the total number of cases reported from 2016 to 2018, during which DENV-2 and DENV-3 gradually expanded (Figure 3A), before cases re-escalated in 2019. Only a single genotype of DENV-2 (DENV-2 cosmopolitan genotype among 99.9% (1822/1823) of DENV-2 study sequences) and two genotypes of DENV-3 (Genotype I; 41.9%; 577/1,327 and genotype III; 58.1%, 800/1,327) contributed to the case load, especially after 2015. DENV-3 genotype I was more common until the Q2-2017 than genotype III, which became more abundant subsequently, and contributed to the escalation of cases in 2019. DENV-2 and DENV-3 expansion after 2015 was also characterized by high intra-genotype diversity, with five genetically distinct clades of cosmopolitan genotype of DENV-2, two clades of genotype I and three clades of genotype III of DENV-3 (Supplementary Figures S2, S3). In DENV-4, genotypes I and II maintained a low profile from 2015 to end-2018 (Figure 4). In overall, genotype II was the most common DENV-4 lineage (90.4%, 273/302), which replaced genotype I in 2019 and became the most dominant genotype among all serotypes in 2021. In contrast to other serotypes, DENV-4 genotype II's dominance in 2021 was associated with the lowest case burden during the study period.

The distribution of common lineages also differed spatially. All common genotypes were mainly detected in Selangor and Johor states (Figure 5). Nevertheless, common genotypes such as DENV-1 genotype I, DENV-2 cosmopolitan genotype and DENV-3 genotype III were widespread throughout the study period (Figure 5). On the other hand, DENV-3 genotype I showed a notable presence in Selangor and Penang in 2016–17 but could not sustain transmission elsewhere in the country.

Bayesian phylogeography analyses demonstrated extensive dispersal of common clades of DENV serotypes between different regions, especially among countries in the South and East Asia as well as between different states in Malaysia (Figures 6-9; Supplementary Table S2). Among 25 significant links identified between Malaysian states and other countries, the highest number of external connections were with Singapore (n = 6), China (n = 4), Bangladesh (n = 3), Myanmar (n = 3) and Japan (n = 3). The highest root state probability values for each common clade showed other regional countries as the most probable ancestral location (Supplementary Table S3), suggesting that the emergence of each common clade in Malaysia was highly likely due to their importation into different states. Moreover, of 50 decisive (BF > 100) diffusion pathways identified (Supplementary Table S2), 36 connections (72%) were between Malaysian states, suggesting a strong and wide network of DENV dispersal within Malaysia. Of them, 34 (94%) connections were within Peninsular Malaysia. Of 143 total number of connections between Malaysian states, Selangor (n = 53) and Johor (n = 13) states demonstrated the highest connectivity with other states (Table 2), highlighting the potential role of these two states as hubs of DENV emergence and spread in Malaysia. This is not surprising because Selangor is an active financial and trade hub connected to other states through a well-developed transportation network. Moreover, Selangor has consistently been the major contributor of dengue burden in the country. The high diversity of DENV in Selangor and Johor also suggested their pivotal role in DENV transmission in Malaysia.

3.4 Proportion of severe infections correlated strongly with the proportion of each genotype within the study cohort

Among 5,471 samples genotyped, 1,662 (30.4%) infections manifested severe dengue. The proportion of severe infections

TABLE 1 Time to most recent common ancestor (tMRCA) analysis of different lineages of DENV serotypes in Malaysia during 2015–2021.

Description	Median node height (95% HPD) in years	Estimated year of emergence[a]	Number of sequences
DENV-1 root ancestor	112 (93.1–136)	1909	NA
Genotype I (clade A)	16 (14.9–18.1)	2005	1,034
Genotype I (clade B)	14 (12.7–15.6)	2007	39
Genotype I (clade C)	15 (13.8–16.5)	2006	327
Genotype I (clade D)	14 (12.1–17.1)	2007	516
Genotype V (clade E)	13 (12.8–14.5)	2008	20
DENV-2 root ancestor	124 (93.8–164.2)	1897	NA
Cosmopolitan (clade 1b)	12 (10.1–13.3)	2009	1,508
Cosmopolitan (clade A)	13 (11.2–15.8)	2008	20
Cosmopolitan (clade B)	13 (10.1–15.3)	2008	22
Cosmopolitan (clade C)	18 (15.6–20.3)	2003	57
Cosmopolitan (clade D)	21 (18.5–23.8)	2000	13
DENV-3 root ancestor	88 (71.3–110.9)	1933	NA
Genotype III (clade A)	11 (9.6–11.9)	2010	484
Genotype III (clade B)	8 (6.6–9.7)	2013	13
Genotype III (clade C)	17 (14.9–18.9)	2004	50
Genotype I (clade D)	16 (13.1–20.6)	2005	257
Genotype I (clade E)	16 (13.0–19.2)	2005	19
DENV-4 root ancestor	150.0 (86.6–287.7)	1871	NA
Genotype I (clade A)	21 (14.0–28.8)	2000	25
Genotype II (clade B)	18 (14.9–21.5)	2003	213
Genotype II (clade C)	13 (9.1–18.4)	2008	19

[a]Estimated year of emergence was calculated by subtracting the node height from 2021, which was the latest temporal data point in the analysis. HPD, highest posterior density.

among the genotyped samples was substantially higher than that among total reported cases during the study period (<1%; Figure 2A) because of the inclusion of only a subset of total infections that sought treatment at 52 sentinel hospitals and clinics throughout the country. Therefore, the sample cohort used for the genotype analysis is likely to have included a relatively higher proportion of symptomatic dengue infections (severe cases and a subset of mild infections) reported during the study period. Those severe infections included 84 fatalities. Overall, the most common genotypes of each serotype contributed the highest proportions of severe infections (Table 3). The number of fatalities also followed a similar trend, with the majority (91.7%; 77/84) being caused by DENV-1 genotype I, DENV-2 cosmopolitan genotype and DENV-3 genotype III, which were in overall the most common genotypes (Table 3). DENV-3 genotype III and DENV-4 genotype II were less likely to cause severe infections compared to DENV-1 genotype I (Table 3). Older individuals were also more likely to suffer from severe infection (Est = 0.14, SE = 0.03, $p < 0.005$), but there was no statistically significant difference between sexes (Est = 0.09, SE = 0.06, $p = 0.162$).

The longitudinal distribution of severe infections due to different genotypes showed that genotype I of DENV-1 and cosmopolitan genotype of DENV-2 collectively contributed the highest number of severe infections in each year from 2015 to 2020 (Figure 10). These two genotypes were the most common during the same period (Figure 4).

4 Discussion

Genomic surveillance forms an integral part of the pathogen surveillance and risk assessment of disease control programs. Genomic data provides valuable insights into the molecular epidemiology of pathogens of interest and thereby facilitates better understanding of the overall epidemiology and control of associated diseases (Pollett et al., 2020). This is especially applicable to fast evolving pathogens, such as RNA viruses, of which the population composition remains highly dynamic over time (Moya et al., 2000; Šimičić and Židovec-Lepej, 2022). The ensuing genetic heterogeneity is known to generate variants of high clinical significance (Šimičić and Židovec-Lepej, 2022). Genomic surveillance gained high importance during the COVID-19 pandemic, emphasizing the need of a global virus surveillance initiative (Li et al., 2022; Hill et al., 2023). Given the ability of leveraging on genome sequencing resources and technical

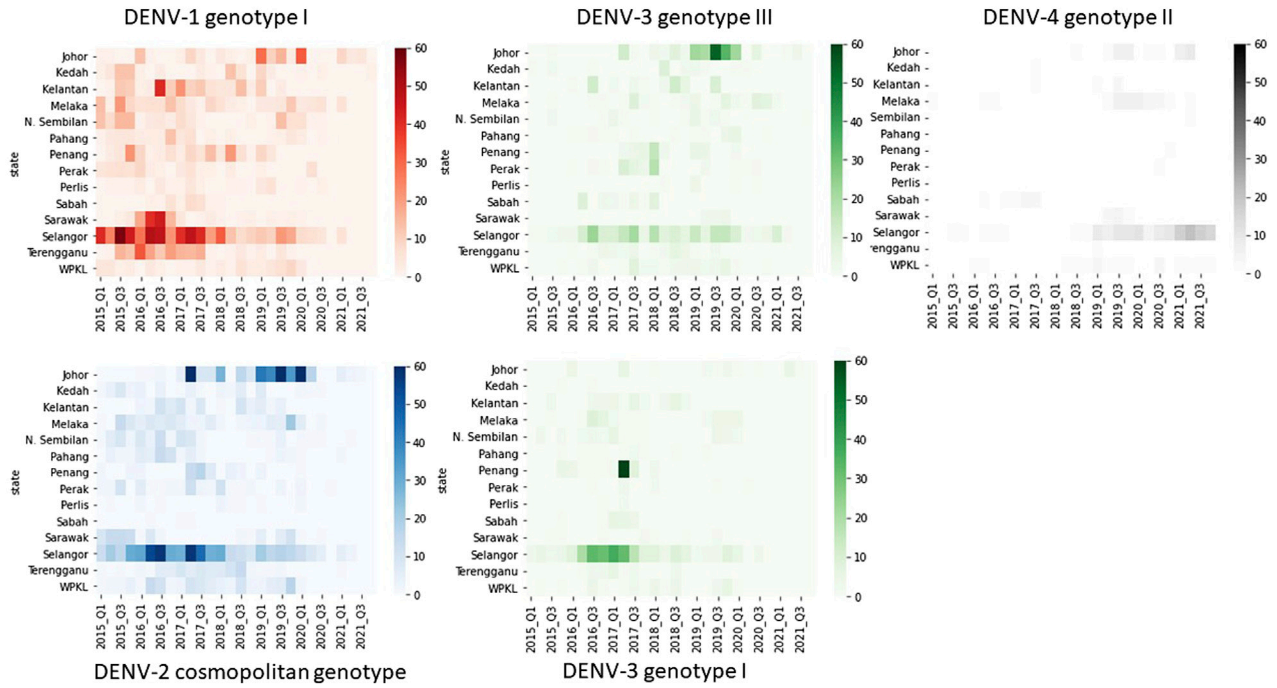


FIGURE 5
Temporal fluctuation of five most common DENV genotypes in different states from 2015 to 2021. The longitudinal pattern of DENV genotypes during the study period in different states of Malaysia was analysed by using a Python 3.6v script. Each rectangle represents data for a 3-month period. Intensity of colours indicates the number of sequences belonging to each genotype as shown in the legend. Q1-Q4 = Quarter one to four.

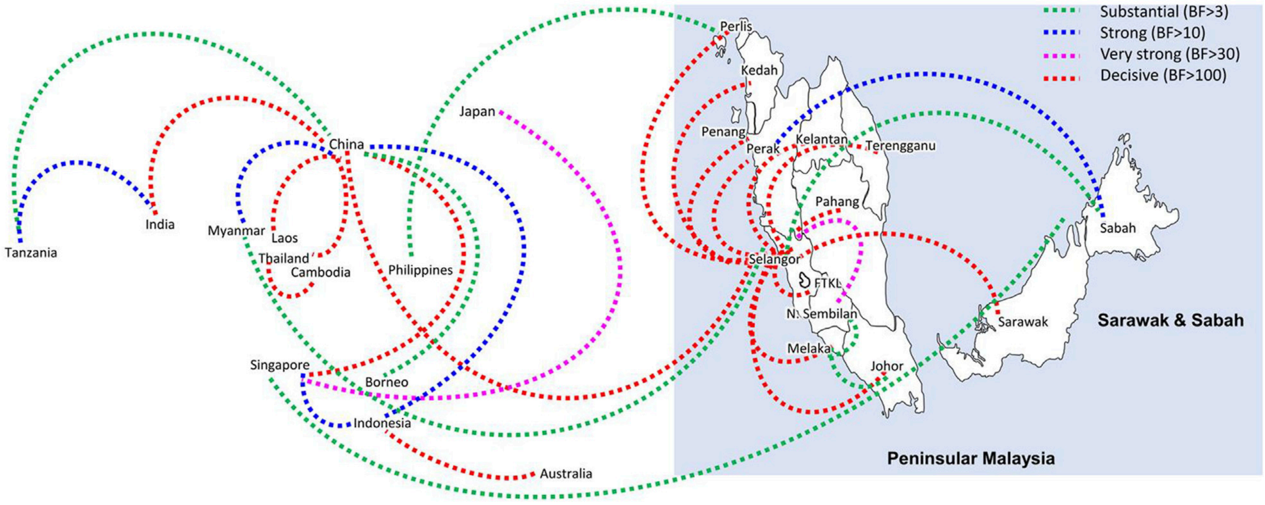


FIGURE 6
The dispersal network of common DENV-1 lineages among different countries and Malaysian states. The dispersal patterns were inferred by using the Bayesian Stochastic Search Variable Selection (BSSVS) procedure in BEAST v1.10.4 (Suchard et al., 2018). Any link supported by Bayes factor (BF) > 3 was considered as significant and only significant links are shown in the figures. The branch colour indicates the BF values as given in the legend (highest in red and lowest in green).

capabilities already established in various countries, World Health Organization launched a 10-year strategy in 2022 to strengthen pathogen genomic surveillance at a global scale (Carter et al., 2022).

Dengue virus is an enveloped RNA virus transmitted by *Aedes* mosquitoes to humans (Roy and Bhattacharjee, 2021). Besides having a high mutation rate characteristic of RNA viruses, DENV's evolution is shaped by its transmission involving a vertebrate and an invertebrate host (Lambrechts and Lequime, 2016; Yu and Cheng, 2022a). Not surprisingly, DENV populations are known to be highly heterogeneous, with constant fluctuations in the proportion of four different serotypes and various genotypes in

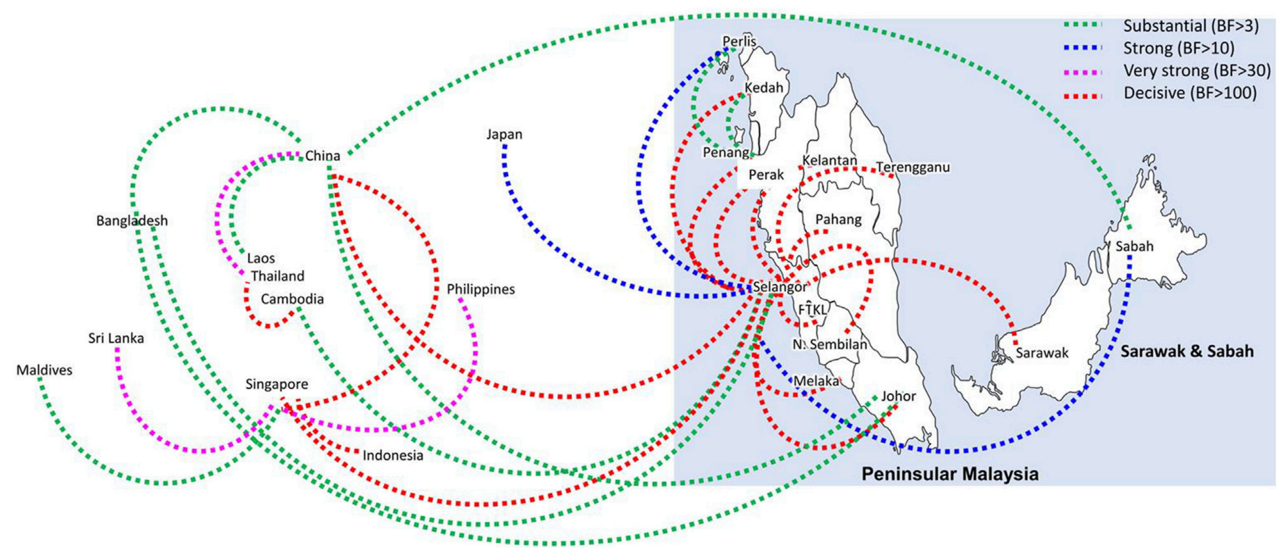


FIGURE 7
The dispersal network of the most common DENV-2 lineage among different countries and Malaysian states. The dispersal patterns were inferred by using the Bayesian Stochastic Search Variable Selection (BSSVS) procedure in BEAST v1.10.4 (Suchard et al., 2018). Any link supported by Bayes factor (BF) > 3 was considered as significant and only significant links are shown in the figures. The branch colour indicates the BF values as given in the legend (highest in red and lowest in green).

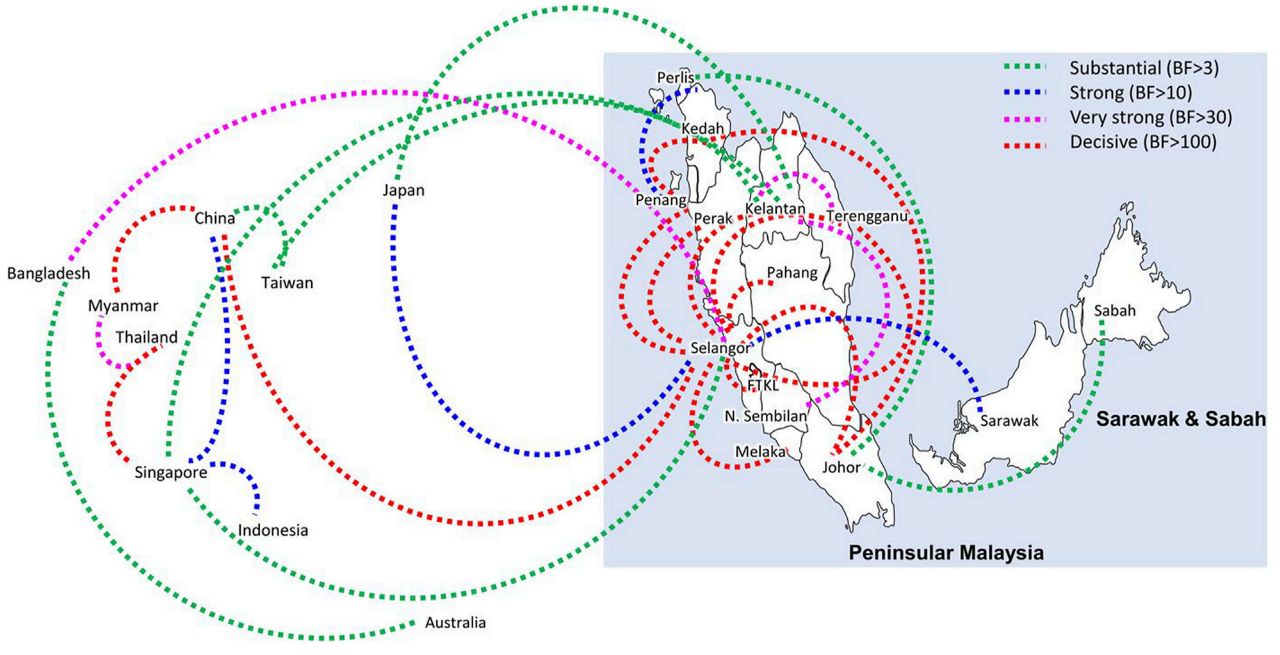


FIGURE 8
The dispersal network of common DENV-3 lineages among different countries and Malaysian states. The dispersal patterns were inferred by using the Bayesian Stochastic Search Variable Selection (BSSVS) procedure in BEAST v1.10.4 (Suchard et al., 2018). Any link supported by Bayes factor (BF) > 3 was considered as significant and only significant links are shown in the figures. The branch colour indicates the BF values as given in the legend (highest in red and lowest in green).

endemic settings (Chen and Vasilakis, 2011). Empirical data has shown that turnover of DENV lineages is often associated with enhanced transmission and outbreaks (Chen et al., 2008; Lambrechts et al., 2012; Tan et al., 2022), emphasizing the importance of genetically characterizing circulating and emerging lineages in affected regions. The present study was therefore aimed at investigating the genetic, evolutionary and dispersal characteristics of DENV serotypes in Malaysia from

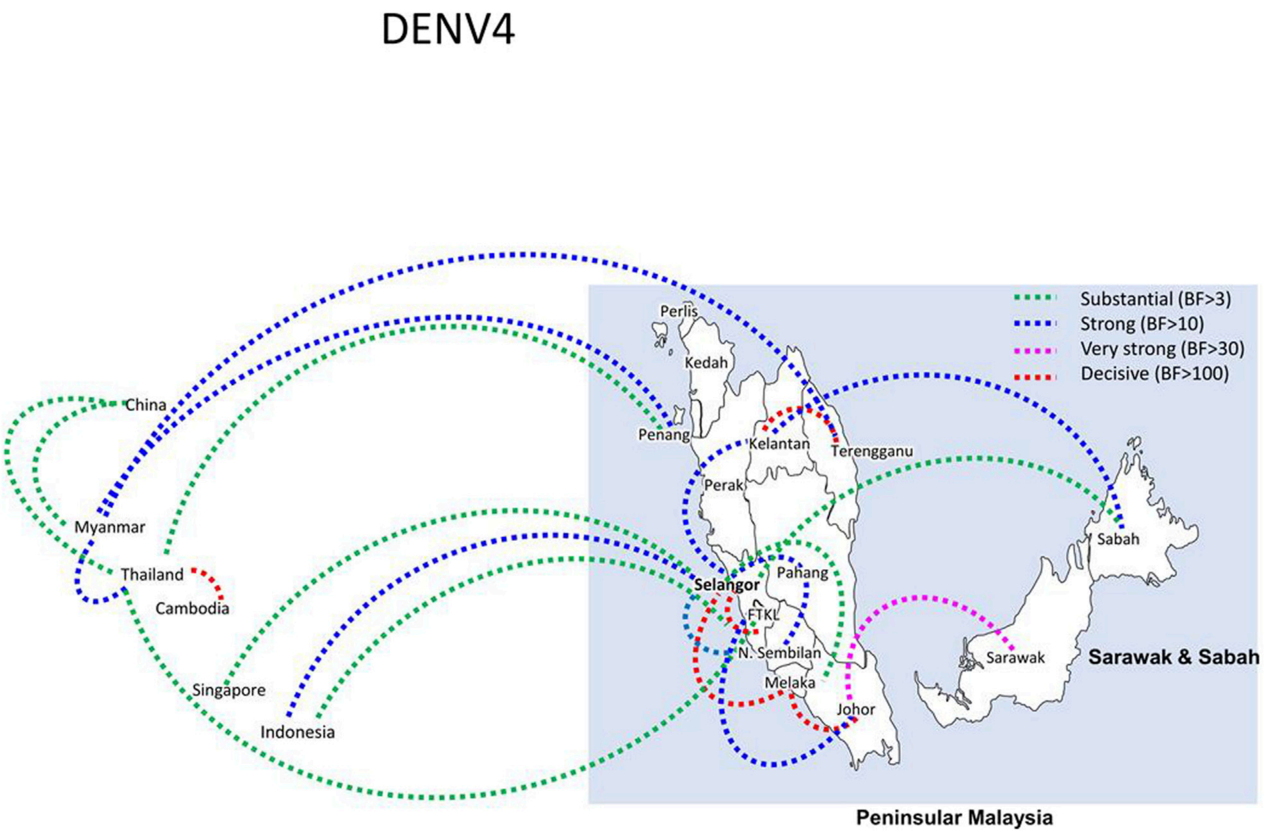


FIGURE 9
The dispersal network of common DENV-3 lineages among different countries and Malaysian states. The dispersal patterns were inferred by using the Bayesian Stochastic Search Variable Selection (BSSVS) procedure in BEAST v1.10.4 (Suchard et al., 2018). Any link supported by Bayes factor (BF) > 3 was considered as significant and only significant links are shown in the figures. The branch colour indicates the BF values as given in the legend (highest in red and lowest in green).

2015 to 2021, during which the overall dengue case burden escalated substantially (Mohd-Zaki et al., 2014; AbuBakar et al., 2022).

The findings revealed that DENV populations in Malaysia during the study period were highly diverse, and the heterogeneity was maintained across space and time through periodic fluctuations of serotypes and lineages (clades) within genotypes. This is a well-described phenomenon in major cities and countries, including Malaysia, where dengue is endemic (Weaver and Vasilakis, 2009; Lee et al., 2010; Rajarethinam et al., 2018; Harapan et al., 2020; AbuBakar et al., 2022; Yu and Cheng, 2022b). All four serotypes co-circulated at any time, confirming the hyperendemic status of dengue in Malaysia (Mia et al., 2013; Mohd-Zaki et al., 2014). Interestingly, the mean nucleotide substitution rate (subs/site/year) of each serotype did not differ substantially and were comparable to empirical data (Chen and Vasilakis, 2011), suggesting that DENV serotypes tend to evolve at a comparable rate, despite temporal fluctuations in their abundance. Besides non-uniformity in sampling that affects the observed heterogeneity in surveillance data, apparently "fixed" rate of evolution among serotypes may also be due to the strong purifying selection acting on DENV populations (Holmes, 2003; Lequime et al., 2016) that may mask the true heterogeneity of natural populations. However, their distribution differed among different states in each year (Mohd-Zaki et al., 2014), suggesting that a single serotype was not uniformly dominant across Malaysia. This observation was further supported by the inability of a single serotype to command a clear dominance across the country during the 7-year study period, except in 2015. Previous studies have shown that dominant serotype switches are associated with dengue outbreaks and such events can be used as early warning to impending outbreaks (Lee et al., 2010). Our findings therefore highlight the importance of having a comprehensive surveillance network with adequate resources in large countries/territories to harness the predictability of outbreaks based on serotype fluctuations at regional level. Moreover, such a network also enhances the ability to improve the spatial resolution of DENV genomic surveillance that is instrumental in monitoring the circulating and newly emerged lineages. As shown in our findings, genetic heterogeneity of DENV demonstrated a wide spatio-temporal spectrum across Malaysia, of which the constituent lineages are likely to have different phenotypes, fitness profiles and epidemic potential. Virus lineages with virulent phenotypes are likely to inflict high hospitalization burden. The study findings did not demonstrate a clear association between the clinical severity and the total number of reported cases. The highest proportion of severe infections was caused by dominant virus lineages. However, none of the virus lineages demonstrated a significantly high risk of causing severe

TABLE 2 Number of diffusion pathways of DENV1-4 with Bayes Factor >3 for each state during 2015–2021.

State	No. of diffusion pathways			
	DENV-1	DENV-2	DENV-3	DENV-4
Johor	2	3	4	4
Kedah	1	2	2	0
Kelantan	1	1	6	3
Malacca	3	1	1	2
Negeri Sembilan	2	1	1	1
Pahang	1	1	1	0
Penang	1	3	3	2
Perak	2	1	1	0
Perlis	2	2	2	0
Sabah	3	2	1	2
Sarawak	1	1	1	1
Selangor	16	16	13	8
Terengganu	1	1	2	2
FTKL	1	3	1	4

FTKL-Federal Territory of Kuala Lumpur.

infections. Instead, DENV-3 genotype III and DENV-4 genotype II were shown to be less likely to cause severe infections compared to the most common virus lineage in the study cohort (DENV-1 genotype I). The findings also showed variable proportions of severe infections when different serotypes were dominant. For example, a higher proportion of severe infections was observed when DENV-1 was dominant in 2015–2016. Interestingly, the percentage of severe infections dropped during the worst dengue outbreak recorded in Malaysia in 2019, when DENV-2 and DENV-3 were dominant. Even though the actual factors contributing to observed differences in severe infection proportions are unclear, study findings highlight the important insights achievable from a surveillance framework integrating clinical and genomic surveillance data. Genetic profiling of circulating DENV populations is also useful when monitoring the efficacy of *Wolbachia-Aedes* replacement strategy that has already been implemented in selected areas in Malaysia (Nazni et al., 2019; Cheong et al., 2023), because the emergence of DENV variants that are able to replicate despite the presence of *Wolbachia* could affect the effectiveness of replacement strategy (Yen and Failloux, 2020).

Having a strong regional genomic surveillance network is further appealing because the dominant genotypes and associated common lineages were distributed through an extensive network, mainly among countries in South and East Asia and different states across Malaysia. The phylogeographic analyses showed direct dispersal of common DENV clades between other countries and Malaysian states, highlighting the importance of understanding virus importation routes into the country. These dispersal routes play a pivotal role in the dissemination of DENV lineages from network hubs to new territories. The findings showed regional countries such as Singapore, China, Bangladesh and Myanmar as the most probable sources of virus introductions and Selangor and Johor as important dissemination hubs of their further spread in Malaysia. This is not surprising because the highest number of dominant lineages were detected in those two states, and they are highly urbanized and densely populated areas with high interstate connectivity to other peninsular regions of Malaysia. Moreover, Johor is the gateway of human movement by ground routes between Malaysia and Singapore, which may be the most potential route of virus sharing between the two countries (Ng et al., 2015). In fact, the most common DENV-2 lineage (Cosmopolitan clade Ib) during the study period is also one of the common lineages circulating in Singapore since 2013 (Hapuarachchi et al., 2016), supporting the high likelihood of cross-border virus sharing. Genomic surveillance at regional level allows the early detection of newly introduced lineages, especially those that have

TABLE 3 The association of DENV genotypes and clinical outcome.

Serotype	Genotype	Mild dengue (n = 3,809)	Severe dengue (n = 1,578)	No. of fatalities (n = 84)	Odds ratio (95% CI)
DENV-1	GI (n = 1,945)	1,309 (67.3%)	611 (31.4%)	25 (1.3%)	Reference
	GII (n = 7)	6 (85.7%)	0	1 (14.3%)	0.31 (0.04–2.56)
	GV (n = 17)	11 (64.7%)	6 (35.3%)	0	1.09 (0.40–2.97)
DENV-2	Asian I (n = 1)	1 (100%)	0	0	NA
	Cosmopolitan (n = 1,822)	1,228 (67.4%)	555 (30.5%)	39 (2.1%)	0.97 (0.85–1.12)
DENV-3	GI (n = 577)	378 (65.5%)	193 (33.4%)	6 (1%)	1.08 (0.89–1.32)
	GIII (n = 800)	629 (78.6%)	158 (19.8%)	13 (1.6%)	0.56 (0.46–0.68)
DENV-4	GI (n = 29)	23 (79.3%)	6 (20.7%)	0	0.52 (0.21–1.28)
	GII (n = 273)	224 (82.1%)	49 (17.9%)	0	0.44 (0.32–0.61)

The percentage values represent the proportion of cases under each clinical category among all samples of respective genotypes. Odds Ratios were derived from logistic regression controlling for age and gender. GI-V, genotypes I to V.

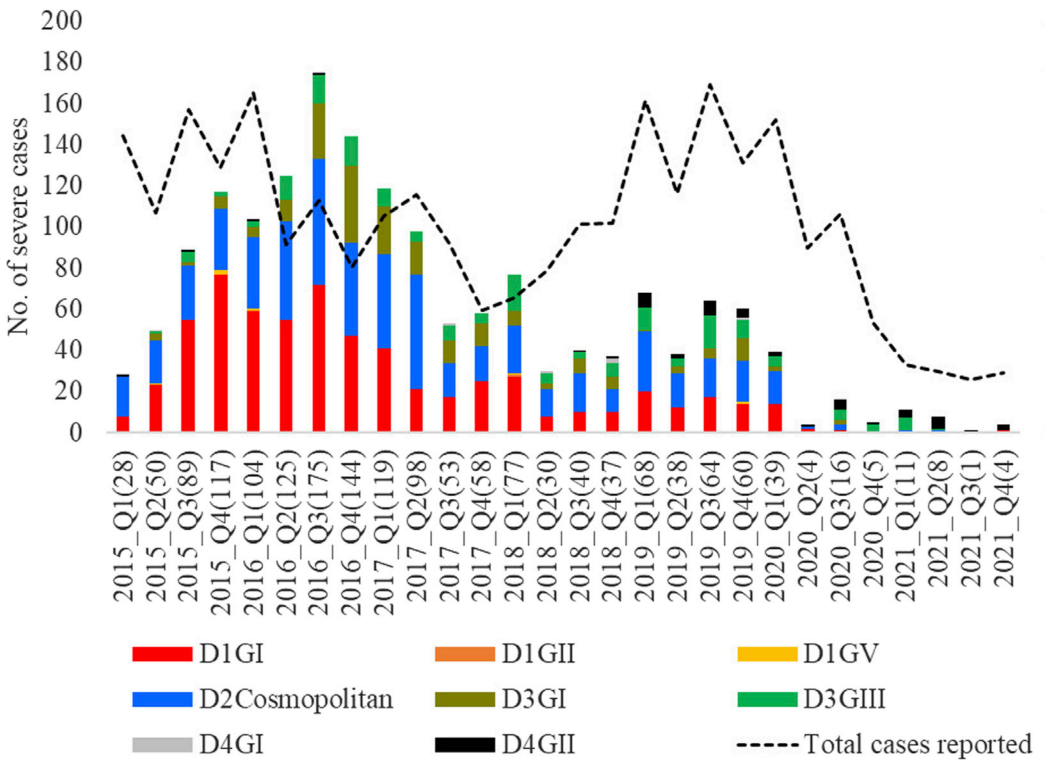


FIGURE 10
Longitudinal distribution of the number of severe infections caused by different genotypes. Severe dengue infections included cases with severe manifestations and dengue-related fatalities. FTKL- = Federal Territory of Kuala Lumpur, Q1-Q4 = Quarter one to four.

demonstrated high epidemic behavior elsewhere, and thereby supports an evidence-based integrated approach to control dengue in large countries, such as Malaysia.

Data availability statement

https://www.ncbi.nlm.nih.gov/genbank/The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/nuccore/pp123917, under the accession numbers PP123917- PP124551.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

YC: Writing–original draft, Methodology, Project administration, Resources, Supervision. SY: Data curation, Formal Analysis, Visualization, Writing–original draft. CC: Project administration, Resources, Supervision, Writing–review and editing. RM: Data curation, Investigation, Writing–review and editing. MW: Data curation, Investigation, Writing–review and editing. NT: Data curation, Investigation, Writing–original draft. RA: Data curation, Investigation, Writing–original draft. MY: Data curation, Investigation, Writing–original draft. RY: Data curation, Investigation, Writing–original draft. CC-C: Data curation, Formal analysis, Writing–original draft. CK: Data curation, Formal Analysis, Visualization, Writing–original draft. LN: Supervision, Writing–review and editing. HH: Conceptualization, Methodology, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article. Not applicable.

Acknowledgments

We would like to thank the Director General of Health, Ministry of Health, Malaysia for his permission to publish this article and the other colleagues of the Ministry of Health, Malaysia involved in the national dengue surveillance program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1368843/full#supplementary-material

References

AbuBakar, S., Puteh, S. E. W., Kastner, R., Oliver, L., Lim, S. H., Hanley, R., et al. (2022). Epidemiology (2012-2019) and costs (2009-2019) of dengue in Malaysia: a systematic literature review. *Int. J. Infect. Dis.* 124, 240–247. doi:10.1016/j.ijid.2022.09.006

Abubakar, S., and Shafee, N. (2002). Outlook of dengue in Malaysia: a century later. *Malays. J. pathology* 24 (1), 23–27.

Ahmad Zaki, R., and Xin, N. Z. (2023). Dengue trend during COVID-19 pandemic in Malaysia. *Asia Pac J. Public Health* 35 (1), 62–64. doi:10.1177/10105395221134655

Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A., and Lemey, P. (2016). SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. *Mol. Biol. Evol.* 33 (8), 2167–2169. doi:10.1093/molbev/msw082

Carter, L. L., Yu, M. A., Sacks, J. A., Barnadas, C., Pereyaslov, D., Cognat, S., et al. (2022). Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032. *Bull. World Health Organ* 100 (4), 239–239a. doi:10.2471/BLT.22.288220

Chee, H. Y., and AbuBaker, S. (2003). Phylogenetic investigation of dengue virus type 2 isolated in Malaysia. *Dengue Bull.* 27, 100–107.

Chen, H. L., Lin, S. R., Liu, H. F., King, C. C., Hsieh, S. C., and Wang, W. K. (2008). Evolution of dengue virus type 2 during two consecutive outbreaks with an increase in severity in southern Taiwan in 2001-2002. *Am. J. Trop. Med. Hyg.* 79 (4), 495–505. doi:10.4269/ajtmh.2008.79.495

Chen, R., and Vasilakis, N. (2011). Dengue--quo tu et *quo vadis*? *Viruses* 3 (9), 1562–1608. doi:10.3390/v3091562

Cheong, Y. L., Nazni, W. A., Lee, H. L., NoorAfizah, A., MohdKhairuddin, I. C., Kamarul, G. M. R., et al. (2023). Spatial distribution and long-term persistence of wolbachia-infected *Aedes aegypti* in the mentari court, Malaysia. *Insects* 14 (4), 373. doi:10.3390/insects14040373

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9 (8), 772. doi:10.1038/nmeth.2109

Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., and Lemey, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos. Trans. R. Soc. Lond B Biol. Sci.* 368 (1614), 20120196. doi:10.1098/rstb.2012.0196

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95–98.

Hapuarachchi, H. C., Koo, C., Rajarethinam, J., Chong, C. S., Lin, C., Yap, G., et al. (2016). Epidemic resurgence of dengue fever in Singapore in 2013-2014: a virological and entomological perspective. *BMC Infect. Dis.* 16, 300. doi:10.1186/s12879-016-1606-z

Harapan, H., Michie, A., Sasmono, R. T., and Imrie, A. (2020). Dengue: a minireview. *Viruses* 12 (8), 829. doi:10.3390/v12080829

Hassan, H., Shohaimi, S., and Hashim, N. R. (2012). Risk mapping of dengue in selangor and Kuala Lumpur, Malaysia. *Geospatial health* 7 (1), 21–25. doi:10.4081/gh.2012.101

Hill, V., Githinji, G., Vogels, C. B. F., Bento, A. I., Chaguza, C., Carrington, C. V. F., et al. (2023). Toward a global virus genomic surveillance network. *Cell Host Microbe* 31 (6), 861–873. doi:10.1016/j.chom.2023.03.003

Holmes, E. C. (2003). Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J. Virol.* 77 (20), 11296–11298. doi:10.1128/jvi.77.20.11296-11298.2003

Iderus, N. H., Singh, S. S. L., Ghazali, S. M., Zulkifli, A. A., Ghazali, N. A. M., Lim, M. C., et al. (2023). The effects of the COVID-19 pandemic on dengue cases in Malaysia. *Front. Public Health* 11, 1213514. doi:10.3389/fpubh.2023.1213514

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Lam, S. K. (1993). Two decades of dengue in Malaysia. *Trop. Med.* 35 (4), 195–200.

Lambrechts, L., Fansiri, T., Pongsiri, A., Thaisomboonsuk, B., Klungthong, C., Richardson, J. H., et al. (2012). Dengue-1 virus clade replacement in Thailand associated with enhanced mosquito transmission. *J. Virol.* 86 (3), 1853–1861. doi:10.1128/JVI.06458-11

Lambrechts, L., and Lequime, S. (2016). Evolutionary dynamics of dengue virus populations within the mosquito vector. *Curr. Opin. Virol.* 21, 47–53. doi:10.1016/j.coviro.2016.07.013

Lee, K. S., Lai, Y. L., Lo, S., Barkham, T., Aw, P., Ooi, P. L., et al. (2010). Dengue virus surveillance for early warning, Singapore. *Emerg. Infect. Dis.* 16 (5), 847–849. doi:10.3201/eid1605.091006

Lee, K. S., Tan, S. S., Chua, R., Tan, L. K., Xu, H., Ng, L. C., et al. (2012). Dengue virus surveillance in Singapore reveals high viral diversity through multiple introductions and *in situ* evolution. *Infect. Genet. Evol.* 12, 77–85. doi:10.1016/j.meegid.2011.10.012

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5 (9), e1000520. doi:10.1371/journal.pcbi.1000520

Lequime, S., Fontaine, A., Ar Gouilh, M., Moltini-Conclois, I., and Lambrechts, L. (2016). Genetic drift, purifying selection and vector genotype shape dengue virus intra-host genetic diversity in mosquitoes. *PLoS Genet.* 12 (6), e1006111. doi:10.1371/journal.pgen.1006111

Li, L., Guo, X., Zhang, X., Zhao, L., Li, L., Wang, Y., et al. (2022). A unified global genotyping framework of dengue virus serotype-1 for a stratified coordinated surveillance strategy of dengue epidemics. *Infect. Dis. Poverty* 11 (1), 107. doi:10.1186/s40249-022-01024-5

Mia, M. S., Begum, R. A., Er, A. C., Abidin, R. D., and Pereira, J. J. (2013). Trends of dengue infections in Malaysia, 2000-2010. *Asian Pac J. Trop. Med.* 6 (6), 462–466. doi:10.1016/S1995-7645(13)60075-9

Mohd-Zaki, A. H., Brett, J., Ismail, E., and L'Azou, M. (2014). Epidemiology of dengue disease in Malaysia (2000-2012): a systematic literature review. *PLoS Negl. Trop. Dis.* 8 (11), e3159. doi:10.1371/journal.pntd.0003159

Moya, A., Elena, S. F., Bracho, A., Miralles, R., and Barrio, E. (2000). The evolution of RNA viruses: a population genetics view. *Proc. Natl. Acad. Sci. U. S. A.* 97 (13), 6967–6973. doi:10.1073/pnas.97.13.6967

Mudin, R. N. (2015). Dengue incidence and the prevention and control program in Malaysia. *Int. Med. J. Malays.* 14 (1), 1–9. doi:10.31436/imjm.v14i1.447

Nazni, W. A., Hoffmann, A. A., NoorAfizah, A., Cheong, Y. L., Mancini, M. V., Golding, N., et al. (2019). Establishment of Wolbachia strain wAlbB in Malaysian populations of *Aedes aegypti* for dengue control. *Curr. Biol.* 29 (24), 4241–4248. doi:10.1016/j.cub.2019.11.007

Ng, L. C., Chem, Y. K., Koo, C., Mudin, R. N. B., Amin, F. M., Lee, K. S., et al. (2015). 2013 dengue outbreaks in Singapore and Malaysia caused by different viral strains. *Am. J. Trop. Med. Hyg.* 92 (6), 1150–1155. doi:10.4269/ajtmh.14-0588

Packierisamy, P. R., Ng, C. W., Dahlui, M., Inbaraj, J., Balan, V. K., Halasa, Y. A., et al. (2015). Cost of dengue vector control activities in Malaysia. *Am. J. Trop. Med. Hyg.* 93 (5), 1020–1027. doi:10.4269/ajtmh.14-0667

Pollett, S., Fauver, J. R., Maljkovic Berry, I., Melendrez, M., Morrison, A., Gillis, L. D., et al. (2020). Genomic epidemiology as a public health tool to combat mosquito-borne virus outbreaks. *J. Infect. Dis.* 221 (Suppl. 3), S308–S318. doi:10.1093/infdis/jiz302

Rajarethinam, J., Ang, L. W., Ong, J., Ycasas, J., Hapuarachchi, H. C., Yap, G., et al. (2018). Dengue in Singapore from 2004 to 2016: cyclical epidemic patterns dominated by serotypes 1 and 2. *Am. J. Trop. Med. Hyg.* 99 (1), 204–210. doi:10.4269/ajtmh.17-0819

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67 (5), 901–904. doi:10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2 (1), vew007. doi:10.1093/ve/vew007

RCoreTeam (2018). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roy, S. K., and Bhattacharjee, S. (2021). Dengue virus: epidemiology, biology, and disease aetiology. *Can. J. Microbiol.* 67 (10), 687–702. doi:10.1139/cjm-2020-0572

Rudnick, A., Tan, E. E., Lucas, J. K., and Omar, M. B. (1965). Mosquito-borne haemorrhagic fever in malaya. *Br. Med. J.* 1 (5445), 1269–1272. doi:10.1136/bmj.1.5445.1269

Šimičić, P., and Židovec-Lepej, S. (2022). A glimpse on the evolution of RNA viruses: implications and lessons from SARS-CoV-2. *Viruses* 15 (1), 1. doi:10.3390/v15010001

Skae, F. M. (1902). Dengue fever in Penang. *Br. Med. J.* 2 (2185), 1581–1582. doi:10.1136/bmj.2.2185.1581-a

Smith, C. E. (1956). Isolation of three strains of type 1 dengue virus from a local outbreak of the disease in Malaya. *J. Hyg.* 54 (4), 569–580. doi:10.1017/s0022172400044843

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4 (1), vey016. doi:10.1093/ve/vey016

Suppiah, J., Ali, E. Z., Mohd Khalid, M. K. N., Mohd Ghazali, S., Tee, K. K., Zulkifli, M. M. S., et al. (2023). Resurgence of dengue virus serotype 4 in Malaysia: a comprehensive clinicodemographic and genomic analysis. *Trop. Med. Infect. Dis.* 8 (8), 409. doi:10.3390/tropicalmed8080409

Sy, A. K., Koo, C., Privaldos, K. J. R., Quinones, M. A. T., Igoy, M. A. U., Villanueva, S., et al. (2023). Genetic diversity and dispersal of DENGUE virus among three main island groups of the Philippines during 2015-2017. *Viruses* 15 (5), 1079. doi:10.3390/v15051079

Tan, C. H., Hapuarachchi, H. C., Tan, L. K., Wong, P. S. J., Li, M. Z. I., Wong, W. Y., et al. (2022). Lineage replacement associated with fitness gain in mammalian cells and *Aedes aegypti*: a catalyst for dengue virus type 2 transmission. *Microorganisms* 10 (6), 1100. doi:10.3390/microorganisms10061100

Teoh, B. T., Sam, S. S., Tan, K. K., Johari, J., Shu, M. H., Danlami, M. B., et al. (2013). Dengue virus type 1 clade replacement in recurring homotypic outbreaks. *BMC Evol. Biol.* 13, 213. doi:10.1186/1471-2148-13-213

Wallace, H. G., Lim, T. W., Rudnick, A., Knudsen, A. B., Cheong, W. H., and Chew, V. (1980). Dengue hemorrhagic fever in Malaysia: the 1973 epidemic. *Southeast Asian J. Trop. Med. public health* 11 (1), 1–13.

Weaver, S. C., and Vasilakis, N. (2009). Molecular evolution of dengue viruses: contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. *Infect. Genet. Evol.* 9 (4), 523–540. doi:10.1016/j.meegid.2009.02.003

WHO (2009). *Dengue: guidelines for diagnosis, treatment, prevention and control*. Geneva: World Health Organization.

Wolf, J., de Souza, A. P., Schardosim, R. F. C., Pille, A., Maccari, J. G., Mutlaq, M. P., et al. (2023). Molecular evolution of dengue virus: a Bayesian approach using 1581 whole-genome sequences from January 1944 to July 2022. *Arch. Virol.* 168 (8), 202. doi:10.1007/s00705-023-05833-3

Yen, P. S., and Failloux, A. B. (2020). A review: wolbachia-based population replacement for mosquito control shares common points with genetically modified control approaches. *Pathogens* 9 (5), 404. doi:10.3390/pathogens9050404

Yu, X., and Cheng, G. (2022a). Adaptive evolution as a driving force of the emergence and Re-emergence of mosquito-borne viral diseases. *Viruses* 14 (2), 435. doi:10.3390/v14020435

Yu, X., and Cheng, G. (2022b). Contribution of phylogenetics to understanding the evolution and epidemiology of dengue virus. *Anim. Model Exp. Med.* 5 (5), 410–417. doi:10.1002/ame2.12283

TYPE Original Research
PUBLISHED 09 July 2024
DOI 10.3389/fmicb.2024.1401259

OPEN ACCESS

EDITED BY
Augusto Simoes-Barbosa,
The University of Auckland, New Zealand

REVIEWED BY
Itziar Chapartegui-González,
Karolinska Institutet (KI), Sweden
Gloria Soberón-Chávez,
National Autonomous University of Mexico,
Mexico
Enoka Corea,
University of Colombo, Sri Lanka

*CORRESPONDENCE
Kevin L. Schully
✉ kevin.schully@nih.gov

RECEIVED 15 March 2024
ACCEPTED 13 June 2024
PUBLISHED 09 July 2024

CITATION
Schully KL, Voegtly LJ, Rice GK, Drumm H, Fitzpatrick MC, Malagon F, Shea A, Dong M, Oduro G, Robberts FJL, Dartey PKA, Owusu-Ofori A, Clark DV, Cer RZ and Bishop-Lilly KA (2024) Phylogenetic and phenotypic characterization of *Burkholderia pseudomallei* isolates from Ghana reveals a novel sequence type and common phenotypes.
Front. Microbiol. 15:1401259.
doi: 10.3389/fmicb.2024.1401259

Phylogenetic and phenotypic characterization of *Burkholderia pseudomallei* isolates from Ghana reveals a novel sequence type and common phenotypes

Kevin L. Schully[1]*, Logan J. Voegtly[2,3], Gregory K. Rice[2,3], Hannah Drumm[2,3], Maren C. Fitzpatrick[2,3], Francisco Malagon[2,3], April Shea[4], Ming Dong[5], George Oduro[6], F. J. Lourens Robberts[7], Paul K. A. Dartey[8], Alex Owusu-Ofori[6,9], Danielle V. Clark[5], Regina Z. Cer[2] and Kimberly A. Bishop-Lilly[2]

[1]Austere Environments Consortium for Enhanced Sepsis Outcomes (ACESO), Biological Defense Research Directorate, Naval Medical Research Command-Frederick, Ft. Detrick, MD, United States, [2]Genomics and Bioinformatics Department, Biological Defense Research Directorate, Naval Medical Research Command-Frederick, Ft. Detrick, MD, United States, [3]Leidos, Reston, VA, United States, [4]National Strategic Research Institute, Omaha, NE, United States, [5]Austere environments Consortium for Enhanced Sepsis Outcomes (ACESO), The Henry M Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD, United States, [6]Komfo Anokye Teaching Hospital, Kumasi, Ghana, [7]Independent Consultant, Stellenbosch, South Africa, [8]CSIR-Crops Research Institute, Kumasi, Ghana, [9]Department of Clinical Microbiology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Melioidosis is a potentially severe disease caused by the gram-negative soil-dwelling bacterium called *Burkholderia pseudomallei*. The true breadth of the distribution of this tropical pathogen is starting to emerge with environmental and clinical isolates frequently characterized in new countries and regions. Even so, isolates, clinical cases, and genetic data from the continent of Africa remain scant. We previously confirmed the presence of *B. pseudomallei* in the environment of Ghana, unmasking a new area of endemicity for this pathogen. Here, we describe the genetic characteristics of isolates obtained from that environmental survey. Twenty-one isolates were subjected to whole genome sequencing and found to represent three discrete sequence types (ST), one of which was novel, and designated ST2058. Phylogenetic analysis places this novel isolate within a *B. pseudomallei* clade that includes genomes derived from the Americas, although it is closely related to a sub-clade that includes isolates from Africa. Importantly, phenotypic characterization demonstrates common features including API 20NE profiles and *B. pseudomallei* CPS to support existing diagnostics, and susceptibility to standard of care antibiotics often used in the clinical management of melioidosis. These findings add to our knowledge about the presence and distribution of *B. pseudomallei* in Africa and represent the first published genomes out of Ghana.

KEYWORDS

***Burkholderia pseudomallei*, melioidosis, Ghana, Africa, environment, phylogenetic, genome, sequence**

Introduction

Burkholderia pseudomallei is a gram-negative soil saprophyte and is the causative agent of the disease known as melioidosis. The bacterium was previously described as being sporadically endemic throughout the tropics with areas of endemic concentration in Thailand and Northern Australia responsible for approximately 2,000 deaths per year. However, in Limmathurotsakul et al. (2016) utilized computer modeling to predict a wider distribution and suggested nearly ubiquitous endemicity throughout the tropical latitudes causing significant morbidity and mortality worldwide (Limmathurotsakul et al., 2016). Consistent with that prediction, *B. pseudomallei* has been identified in clinical and environmental specimens in areas not previously known to be endemic, including South Asia (Mukhopadhyay et al., 2018; Jayasinghearachchi et al., 2022), the Caribbean (Hall et al., 2019; Stone et al., 2020), and North America (Salam et al., 2011; Morosini et al., 2013; Torres, 2023).

Even as *B. pseudomallei* was discovered to have a broader geographic distribution around the world, knowledge of the epidemiology and distribution of *B. pseudomallei* in Africa remained relegated to a small number of sporadic cases, mostly exported from Africa, and anecdotal accounts (Wall et al., 1985; Cuadros et al., 2011; Sarovich et al., 2016; Steinmetz et al., 2018; Mabayoje et al., 2022). Efforts to remedy that knowledge gap have recently increased through campaigns to raise awareness, increase diagnostic capabilities, and identify *B. pseudomallei* in the environment (Steinmetz et al., 2018; Birnie et al., 2019, 2022; Savelkoel et al., 2023). We recently conducted an environmental survey of five rice paddies in South-central Ghana and confirmed the presence of *B. pseudomallei* in Ghana through standard culture and biochemical approaches, and found that these isolates were susceptible to standard antimicrobial therapies for melioidosis (Oduro et al., 2022). Although the epidemiological picture of *B. pseudomallei* in Africa is starting to emerge, genetic data remains scant with only 69 of the 6,817 isolates in PubMLST associated with Africa (assessed on December 26, 2023). Of these, only nine genome assemblies are available; notably two (id: 4412 and 4413) originated from Burkina Faso, the northern neighbor of Ghana. The other seven assemblies originated from Nigeria. GenBank has 144 complete *B. pseudomallei* (taxonomy ID: 28450) genomes available but none are from Africa (Supplementary Table 1).

B. pseudomallei has a large, dynamic, and genetically diverse genome owing to horizontal gene transfer and site-specific recombination occurring at integration hotspots located throughout its genome (Tuanyok et al., 2008). This genomic plasticity makes genetic characterization a complicated but essential component of the description of potentially novel strains. Multi-locus sequence typing (MLST) is a commonly used method for characterizing the epidemiology of *B. pseudomallei*, the diversity within environmental samples, and origin of clinical specimens (Godoy et al., 2003; Dale et al., 2011; Roe et al., 2022). However, the granularity and resolution that whole genome sequencing (WGS) provides for genomic characterization remains the most accurate method, particularly for molecular epidemiological studies related to outbreak surveillance and for investigations of biogeography and diversity. In our previous study, we identified Ghana as an endemic area for *B. pseudomallei* and described the phenotypic characteristics of isolates obtained from one of five sites (Figure 1, site E) of this environmental survey (Oduro et al., 2022). In this follow-up study, we describe the genetic characteristics of isolates obtained from 5 sampling sites (Figure 1) from which 21 isolates were found to represent 3 discrete sequence types (ST), 1 of which was novel. These findings add to our knowledge about the presence and distribution of *B. pseudomallei* in Africa and represent the first published genomes that originated in Ghana.

Materials and methods

Environmental sample collection

Soil samples were collected in January 2021 from 100 points within each of five sites, each 35 m^2, around the Ashanti region of Ghana as previously described and shown in Figure 1 (Oduro et al., 2022). Sampling environment and sampling site details are presented in Figure 2F and Supplementary Table 2, respectively. Soils were enriched for *B. pseudomallei* using a combination of methods previously described. The enriched samples were preserved by combining 1.2 ml of the resulting cultures with 0.3 ml of 80% sterile glycerol and frozen at −80°C for shipment and storage. Frozen cultures were thawed and used directly to prepare crude DNA extracts for screening by PCR. PCR-positive samples were colony purified on Ashdown's agar and pure cultures of suspected *B. pseudomallei* were utilized for DNA isolation and phenotypic characterization.

PCR and culturing

Briefly, crude DNA extracts were prepared by lysing 100 μl of frozen stocks at 95°C for ten minutes. Cell debris was pelleted by centrifugation and 2 μl of the resulting supernatant was used as template in the PCR reaction as previously described (Novak et al., 2006). Each extract was analyzed in duplicate, and each reaction plate contained positive controls including *B. pseudomallei* strain Bp82 crude extract and purified Bp82 genomic DNA, as well as negative controls including *B. thailandensis* strain E264 crude extract and water. To identify extracts containing *B. pseudomallei* genomic DNA, cutoff values were generated by plotting the results of true positive samples versus true negative samples using Graphpad Prism (version 9.0, Boston, Massachusetts USA).[1] Frozen stocks from sites determined to be positive by PCR were used to inoculate Ashdown's agar plates which were incubated at 35°C for 96 h. Colonies presenting a morphology consistent with *B. pseudomallei* (i.e., flat, wrinkled, purple colonies) were selected for genetic characterization.

Genomic DNA isolation

Suspected *B. pseudomallei* colonies were inoculated into 5 ml tryptic soy broth (TSB) and incubated at 37°C overnight while shaking. The following day, 1 ml was withdrawn, the cells were

[1] www.graphpad.com

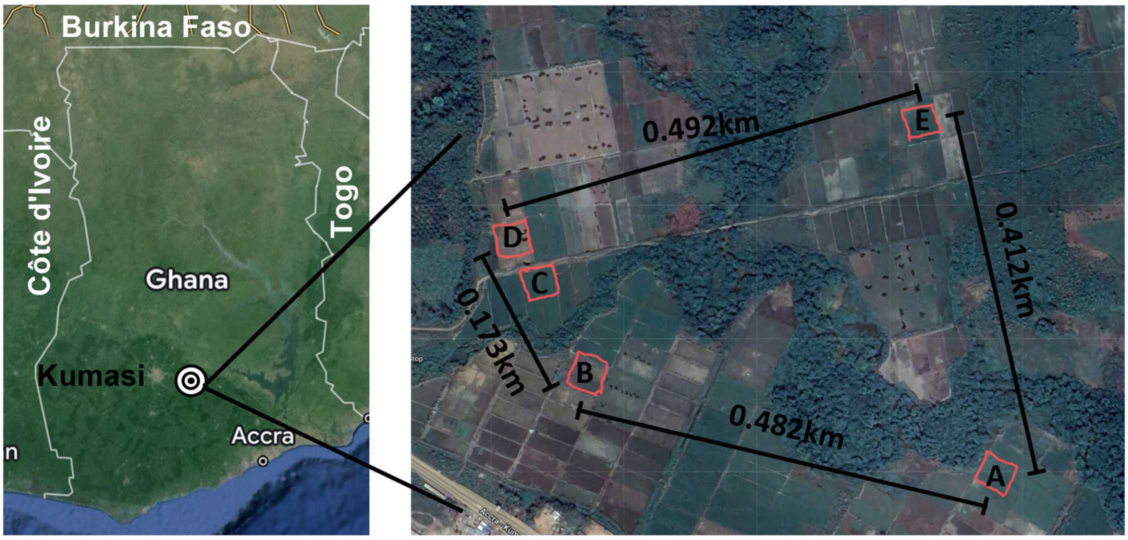


FIGURE 1
Sampling area. **Left:** Map of Ghana with major reference points included and the sampling area indicated with concentric circles. **Right:** Each of the five sites is indicated at scale with the distance between each noted on the map. A mobile phone field measurement tool, GPS Fields Area Measure (FARMIS; Lithuania), was used with an Apple iPhone X (Apple Inc.; Cupertino, CA, USA) to record the Global Positioning Coordinates (GPS Coordinates., 2023) of each sampling site by tracing the coordinates while walking around the demarcated perimeters of each sampling site. The resulting field shape coordinates were exported as KML files. The distance between each site was determined using the Distance Calculator.

pelleted, and genomic DNA was isolated using the MasterPure Complete DNA and RNA Purification Kit (LGC Biosearch Technologies; Middleton, WI) according to the manufacturer's instructions. Purified genomic DNA was quantified using Qubit dsDNA BR assays (ThermoFisher Scientific; Waltham, MA, USA) and a Qubit 4 fluorometer. DNA integrity was evaluated by electrophoresis using Genomic DNA ScreenTape, genomic reagents and a TapeStation 4150 (Agilent Technologies; Santa Clara, CA, USA).

Sample selection for sequencing

Each sampling site of 35 m^2 was divided into a 10 × 10 grid of 3.5 m^2 (Oduro et al., 2022). To select colonies that represented each site, we divided each site into four quadrants of 25 sampling points. Due to the uneven distribution of PCR-positive sites in sampling site C (Figure 2), the upper left quadrant was reduced to 20 points and the lower left quadrant was expanded to 30 points (Figure 2). The highest quality DNA sample from each quadrant was selected for sequencing. If a given quadrant had only one sample, then that sample was chosen regardless of DNA concentration. Specific locations are indicated with a star in the grids of Figure 2.

Sequencing

Short-reads shotgun libraries were prepared using NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England Biolabs; Ipswich, MA, USA) following the manufacturer's instructions. Briefly, 26 μl of DNA at 2–4 ng/μl was first fragmented enzymatically via 20 min incubation at 37°C. Next, hairpin sequencing adaptors, containing 5′-dT overhangs and a U ribonucleotide in the hairpin loop, were added to the fragmented DNA by ligation, and subsequently cleaved at the U sites. The libraries were then amplified and indexed by PCR using NEBNext Unique Dual Indexes. Prior to sequencing, the libraries were evaluated for quality using Agilent D1000 kit (Agilent Technologies; Santa Clara, CA, USA). The libraries that passed quality control were then quantified using Qubit dsDNA BR assay (ThermoFisher Scientific; Waltham, MA), pooled, and sequenced using a NovaSeq 6000 S4 Reagent Kit, v1.5 300 cycles, and a NovaSeq 6000 sequencer (Illumina; San Diego, CA, USA). Long-reads libraries were prepared using Ligation kits with native barcodes (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions. Briefly, 8 μl of DNA at ∼50 ng/μl was first polished for ligation using NEBNext FFPE DNA Repair Mix. The end-prepped DNA was then indexed by incubation with ONT Native Barcodes and Blunt/TA Ligase Master Mix (NEB). The indexed libraries were then pooled and ligated to ONT AMXII sequencing adaptors using NEBNext Quick T4 DNA ligase. Free adapters and short library fragments were eliminated by cleaning with AMPure beads (Beckman Coulter; Brea, CA) and ONT Long Fragment Buffer. Finally, the libraries were sequenced using a MinION_Flow Cell (R9.4.1) and a MinION-MK1C sequencer (Oxford Nanopore Technologies, Oxford, UK).

Quality control and *de novo* assembly of sequencing data

The NovaSeq data were processed using MetaDetector as previously described (Adhikari et al., 2024). Briefly, via this pipeline, reads were trimmed and filtered using BBDuk (v38.96) (Bushnell, 2014) and the resulting data were assembled using SPAdes (v.3.15.3) (Bankevich et al., 2012). The reads were mapped

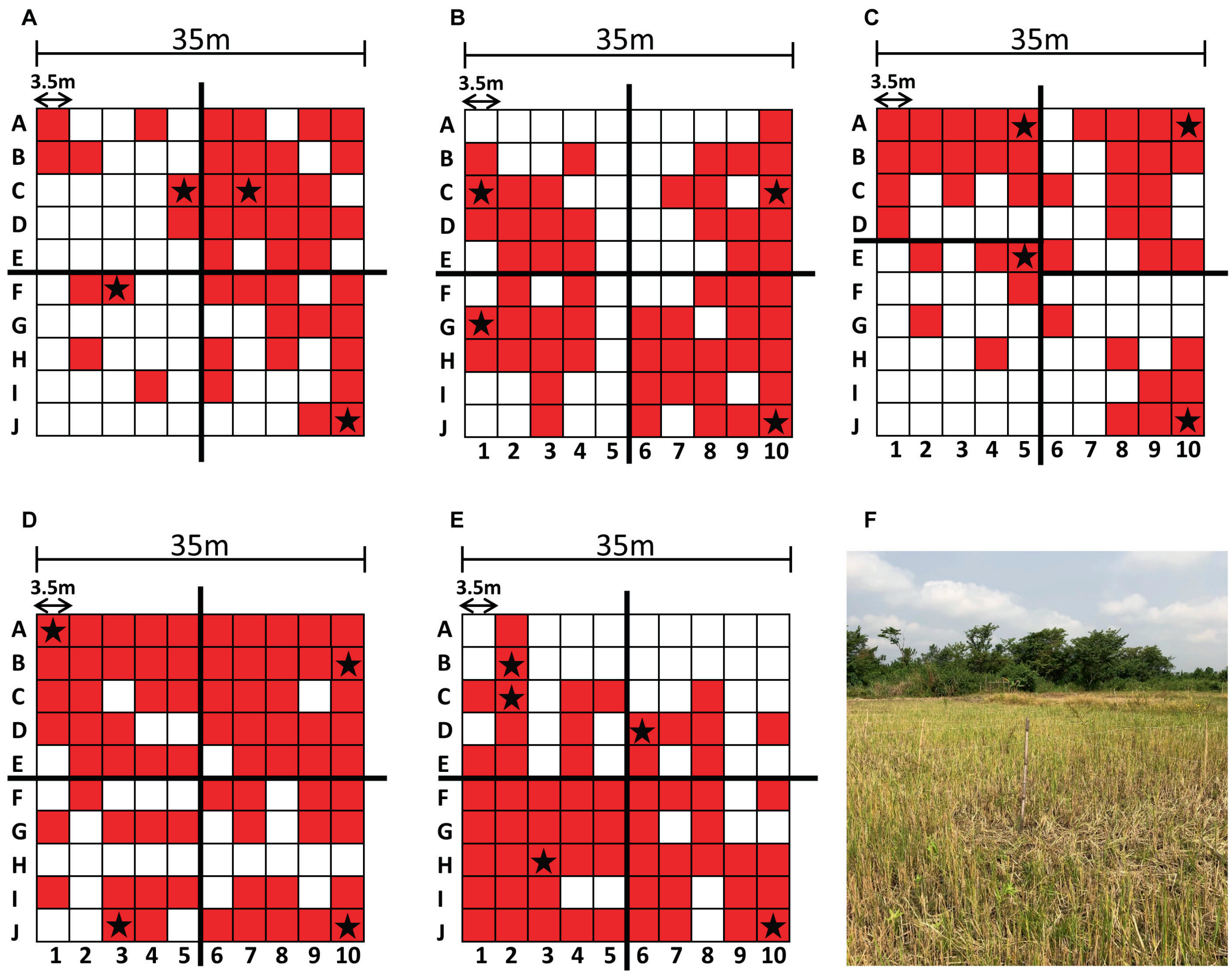


FIGURE 2
Sampling results and environment. Each sampling grid is depicted with the PCR-positive sampling points shaded in red and the sequenced genomes indicated with a star **(A–E)**. The augmented quadrants are shown in **(C)**. Sampling site **(A)** is shown as a representative example of the sampling environment of each sampling site **(F)**.

TABLE 1 Potential cutoff values indicating PCR-confirmed *B. pseudomallei*.

Ct Cutoff	Sensitivity %	95% CI	Specificity %	95% CI
> 24.55	100.0	77.19 to 100.0%	91.67	64.61 to 99.57%
> 30.47	**100.0**	**77.19 to 100.0%**	**100.0**	**75.75 to 100.0%**
> 37.28	92.31	66.69 to 99.61%	100.0	75.75 to 100.0%

The sensitivity and specificity are provided for three potential cutoff Ct values, as well as the confidence interval (CI) for each of those values. The Ct value for 100% sensitivity and 100% specificity is presented in bold.

back to the contigs using BBMap (v38.96) (Bushnell, 2014), and the reads and contigs were classified using DIAMOND BLAST against NCBI's non-redundant protein database (nr accessed February 10, 2023) (Buchfink et al., 2015). The ONT reads were assembled using dragonflye (v1.0.14) (Kolmogorov et al., 2019). In addition to these two methods, trimmed and filtered NovaSeq reads were further subsampled to 20 million (M) reads using seqtk (v1.3-r106) and combined with ONT reads for a hybrid assembly using UniCycler (v0.5.0) (Wick et al., 2017).

Genome closure

The most cohesive UniCycler assemblies were selected from each site (except for SiteE where Drangonflye assembly was used) and manually extracted using Bandage (v0.9.0) (Wick et al., 2015). The extracted genomes were validated using Qiagen CLC Genomics Workbench (v23.0.2)[2] by mapping the subsampled 20M NovaSeq reads and Oxford Nanopore Technologies (ONT) reads against the draft genome and using the *Analyze Contigs* function to identify problematic regions and *Basic Variant Detection* to identify regions with differences in the reads and contigs (CLC Microbial Genomics Module 23.0). The problematic regions and differences were then manually resolved. This process was repeated until there were

[2] https://digitalinsights.qiagen.com/

TABLE 2 Detailed information on MLST sequence types of the *B. pseudomallei* isolates.

Sample point	Strain name	Chromosome 1 length (bp)	Chromosome 2 length (bp)	lipA	gltB	lepA	gmhD	narK	ace	ndh	MLST sequence type	Core genome sequence type	cgMLST kmer fraction
SiteA-3F	GHA3F			5	1	2	3	1	1	1	ST930		
SiteA-5C	GHA5C			5	1	2	3	1	1	1	ST930		
SiteA-7C	GHA7C			5	1	2	3	1	1	1	ST930		
SiteA-10J	GHA10J	4,034,701	3,217,548	5	1	2	3	1	1	1	ST930	cgST1	0.48
SiteB-1C	GHB1C			5	1	2	3	1	1	1	ST930		
SiteB-1G	GHB1G			5	1	2	3	1	1	1	ST930		
SiteB-10C	GHB10C			5	1	2	3	1	1	1	ST930		
SiteB-10J	GHB10J			5	1	2	3	1	1	1	ST930		
SiteC-5A	GHC5A			6	1	2	2	1	1	1	ST1749		
SiteC-5E	GHC5E	4,061,312	3,203,862	6	1	2	2	1	1	1	ST1749	cgST812	0.52
SiteC-10A	GHC10A			6	1	2	2	1	1	1	ST1749		
SiteC-10J	GHC10J			6	1	2	2	1	1	1	ST1749		
SiteD-1A	GHD1A	4,051,091	3,134,872	6	1	2	13	2	1	3	ST2058	cgST571	0.46
SiteD-3J	GHD3J			6	1	2	13	2	1	3	ST2058		
SiteD-10B	GHD10B			6	1	2	13	2	1	3	ST2058		
SiteD-10J	GHD10J	4,051,098	3,134,866	6	1	2	13	2	1	3	ST2058	cgST571	0.46
SiteE-2B	GHE2B			5	1	2	3	1	1	1	ST930		
SiteE-2C	GHE2C			5	1	2	3	1	1	1	ST930		
SiteE-3H	GHE3H			5	1	2	3	1	1	1	ST930		
SiteE-6D	GHE6D	4,046,104	3,215,367	5	1	2	3	1	1	1	ST930	cgST1	0.47
SiteE-10J	GHE10J			5	1	2	3	1	1	1	ST930		

no problematic regions or differences identified, resulting in a high-quality draft genome.

Genome annotation and multilocus sequencing typing (MLST)

Genome annotation was performed using EDGE Bioinformatics (v2.4.0 BDRD 2023FEB09) (Li et al., 2017) with Prokka (v1.14.0) (Seemann, 2014). Virulence factors were identified using EDGE with ShortBRED (v0.9.4M) (Kaminski et al., 2015), antibiotic resistance genes were identified using EDGE with Resistance Gene Identifier (RGI v5.0.1; database v3.0.7) (Jia et al., 2017), and insertion sequences were identified using BLAST against the ISFinder database (accessed 2023-05-04) (Siguier et al., 2006). The MLST database for *B. pseudomallei* was downloaded via CLC on March 28, 2023, and May 04, 2023. CLC Type with MLST Scheme function was used with the SPAdes contigs, and the high-quality draft genomes, to identify sequence types for each of the samples. CLC Genomics Workbench was used to pull the core genome multilocus sequence typing (cgMLST) database and perform the cgMLST to determine a possible cgST. The cgMLST database has 4,071 alleles and requires a minimum fraction of 0.50 to have confidence in typing.

Comparative genomic analyses

A total of 1,777 complete assemblies were downloaded from GenBank and nine assemblies associated with Africa were downloaded from PubMLST on May 04, 2023. Snippy (v4.6.0) (Seemann, 2015) was used to generate an alignment of the core single nucleotide polymorphism (SNP) genomes using *B. pseudomallei* Mahidol-1106a (Assembly Accession GCA_000756125.1) from Thailand as the reference. Using the core SNP alignment, IQ-Tree (v1.6.10) (Nguyen et al., 2015) was used to generate a Maximum Likelihood tree with automatic model testing using TVM+F+ASC+G4 and 1,000 bootstrap and BEAST (v1.10.4) was used to generate a Bayesian tree using HKY model and 1M MCMC chains. Mauve (v2.4.0) (Darling et al., 2004) was used to perform a whole genome alignment of the three high-quality draft genomes and reference *B. pseudomallei* Mahidol-1106a.

Phenotypic characterization of *Burkholderia pseudomallei* isolates

Phenotypic characterization such as enzymatic activity and carbohydrate utilization were evaluated using the Analytical Profile Index system, specifically API 20NE (BioMerieux; Cambridge, MA, USA). Antibiotic sensitivity testing (AST) using the disk diffusion method and detection of *B. pseudomallei* capsular polysaccharide by the Active Melioidosis Detect (InBios International Inc, Seatle, WA, USA) lateral flow immunoassay were conducted as previously described (Oduro et al., 2022).

Results

Screening of environmental samples for *B. pseudomallei*

Soil cultures were generated from a soil sampling expedition in Ghana using the selective enrichment process described by Trinh and colleagues (Trinh et al., 2019). We first sought to screen these

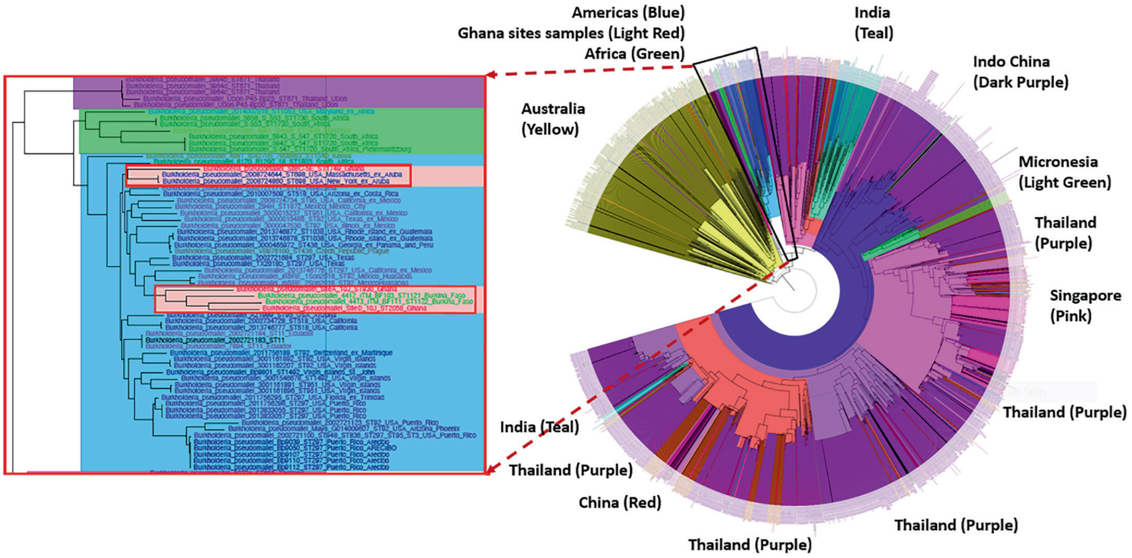


FIGURE 3
Phylogenetic analyses of the *Burkholderia pseudomallei* genomes from five different sites. **Left:** The genomes from Ghana grouped within a clade associated with isolates from the Americas. **Right:** The isolates from Site A (ST930) and the new ST2058 from Site D grouped within a sub-clade that includes isolates from Ghana-neighboring Burkina Faso. ST1749 from Site C grouped closely with isolates from Aruba.

cultures to down-select from our 500 soil samples, using PCR-specific for *B. pseudomallei*. Real time PCR targeting *orf2* of the *B. pseudomallei* Type III Secretion System is considered 100% specific for *B. pseudomallei* because this region is not present in any near neighbor species (Novak et al., 2006). Ct values of true positive results derived from crude extracts of *B. pseudomallei* strain Bp82, prepared as described in the "Materials and methods" section, were plotted against the negative values to generate a Receiver Operator Characteristic (ROC) curve to identify a cutoff value. As depicted in Supplementary Figure 1, the area under the curve (AUC) was 1.0 [95% Confidence Interval (CI) 1.0 to 1.0]. The ROC curve identified a PCR cutoff value for 100% sensitivity and 100% specificity (Table 1). By applying a cutoff value of >30.47 to DNA extracts from each of the 100 points across each site, we found the positive rates as follows: Site A, 45%; Site B, 56%; Site C, 43%; Site D, 70% and Site E, 60%. These results confirm our previous observation that *B. pseudomallei* is ubiquitous throughout the Ashanti region of Ghana, although the distribution is not even (Figure 2 and Supplementary Table 2) (Oduro et al., 2022).

Sequencing and *de novo* assembly results

We performed whole-genome sequencing using both short read and long read platforms on 21 isolates from the five sampling sites around the Ashanti Region of Ghana (Figures 1, 2) with *de novo* assembly of the resulting sequencing data (Supplementary Table 3). Each genome from the same site was found to be very similar (i.e., a minimum of 97.92% nucleotide identity percentage); therefore, one representative genome from each site was deemed to be sufficient rather than producing high-quality drafts of duplicate genomes. To that end, five genomes from four sampling sites including one each from SiteA_10J, SiteC_5E, SiteD_1A, SiteD_10J, and SiteE_6D were manually closed to high quality draft genome status (Supplementary Table 4). An isolate from Site B was not chosen due to a high degree of sequence similarity to isolates from Site A.

MLST sequence typing and phylogenetic analyses

Isolates from three of the sampling sites represented previously identified MLST Sequence Types (ST). ST930 was found in site A and site E and is represented by isolates found in four countries including the soil of Nigeria, located to the East of Ghana (Savelkoel et al., 2023). ST1749 from sampling site C is represented by a single isolate from Mexico.[3] The genome of isolates from sampling site D was found to have a novel sequence type, now designated ST2058 (Table 2).

Phylogenetic analyses showed that the genomes from Ghana grouped within a clade associated with isolates from the Americas. The isolates from site A (ST930) and the two isolates designated with the new strain type ST2058 grouped within a sub-clade that

TABLE 3 Phenotypic characterization and results.

	B. pseudomallei API profile 1156577																					
	NO3	TRP	GLU	ADH	URE	ESC	GEL	PNPG	GLU	ARA	MNE	MAN	NAG	MAL	GNT	CAP	ADI	MLT	CIT	PAC	OX	CPS
GHC5E	+	–	–	+	–	–	+	–	+	–	+	+	+	–	+	+	+	+	+	+	+	+
GHD1A	+	–	–	+	–	–	+	–	+	–	+	+	+	–	+	+	+	+	+	+	+	+
GHD10J	+	–	–	+	–	–	+	–	+	–	+	+	+	–	+	+	+	+	+	+	+	+
GHE2C	+	–	–	+	–	–	+	–	+	–	+	+	+	–	+	+	+	+	+	+	+	+

Bacterial identification by API 20NE. For each strain, the API profiles are shown alongside their reaction to *B. pseudomallei* capsular polysaccharide (CPS) LFI. NO_3, potassium nitrate reduction; TRP, L-tryptophane (indole production); GLU, D-glucose fermentation; ADH, L-arginine DiHydrolase; URE, urea (urease production); ESC, esculin hydrolysis; GEL, gelatin hydrolysis; PNPG, 4-nitrophenyl-βD-galactopyranoside (β-galactosidase production); GLU, D-glucose assimilation; ARA, L-arabinose assimilation; MNE, D-mannose assimilation; MAN, D-mannitol assimilation; NAG, N-acetyl-glucosamine assimilation; MAL, D-maltose assimilation; GNT, potassium gluconate assimilation; CAP, capric acid assimilation; ADI, adipic acid assimilation; MLT, malic acid assimilation; CIT, trisodium citrate assimilation; PAC, phenylacetic acid assimilation; OX, cytochrome oxidase. + indicates a positive test.

[3] https://pubmlst.org

TABLE 4 Antibiotic susceptibility profiles.

Organism	Strain	Zone of inhibition (mm) for antimicrobial agent					
		CAZ (30 μ g)	IPM (10 μ g)	MEM (10 μ g)	DO (30 μ g)	AMC (20/10 μ g)	SXT (1.25/23.75 μ g)
B. pseudomallei	GHC5E	29	28	25	27	27	31
	GHD1A	29	37	26	29	27	31
	GHD10J	28	35	26	26	28	31
	GHE2C	30	35	27	29	28	31

Antibiotic susceptibility profiles for isolates from Ghana. Each disk and its content (in parentheses) are provided, as are the zones of inhibition for amoxicillin-clavulanic acid (AMC), ceftazidime (CAZ), doxycycline (DO), imipenem (IPM), meropenem (MEM), and trimethoprim-sulfamethoxazole (SXT).

includes isolates from Ghana-neighboring Burkina Faso, whereas the isolate with strain type ST1749 from Site C grouped closely with isolates from Aruba (Figure 3). We also observed that all genomes obtained from these sites contained the *Yersinia*-like fimbrial (YLF) (also referred to as fimbria/pilus outer membrane usher protein) gene. The presence of YLF is in agreement with a 2014 study by Gee et al. (2017) that established that YLF gene is associated mainly with isolates from locations other than Australia.

Phenotypic characterization

We conducted phenotypic characterization, including enzymatic activity, carbon utilization, capsular polysaccharide (CPS) detection and AST analyses of representative isolates of each of the STs including Site C, point 5E (C5E) representing ST1749, Site E point 2C (E2C) representing ST930 found in sites A and E. We selected two isolates representing our new ST from sampling site D (D1A and D10J). These isolates were from the farthest points on the sampling grid (Figure 2) and were selected due to minor differences in their genome sequences that could have translated into phenotypic variations. Each isolate was oxidase positive, was CPS positive, and produced an API 20NE profile of 1156577, one of the most common *B. pseudomallei* profiles (Table 3) (Amornchai et al., 2007). Each isolate was also susceptible to common antibiotics used in the management of melioidosis (Table 4).

Discussion

Here, we present the first genotypic and second phenotypic analyses of *B. pseudomallei* isolates obtained from the soil of Ghana. Out of 21 genomes sequenced, we identified three discrete sequence types including a unique ST, hereby designated ST2058. SNP-based phylogenetic analyses of the core genomes show that isolates from Ghana cluster within the American Clade, and an African sub-clade, of known *B. pseudomallei* genomes. These results are consistent with other studies and a recent hypothesis proposing an African origin of *B. pseudomallei* in the Americas, potentially seeded by the Atlantic slave trade (Chewapreecha et al., 2017). Additionally, the YLF gene cluster was universally present in all of the sequences from this study. The YLF cluster is typically found in isolates from Southeast Asia, as opposed to the *Burkholderia thailandensis*-like flagellum and chemotaxis biosynthesis (BTFC) gene cluster, which are often found in the genomes of isolates from Australia (Tuanyok et al., 2008). These genomes can now be added to the 20 *B. pseudomallei* genomes from the Americas confirmed to have YLF by BLAST searches (Supplementary Table 5) in future *B. pseudomallei* phylogeographic studies (Gee et al., 2017). Because YLF is overrepresented in clinical isolates, it is presumed to be a virulence factor, although its significance in disease is unknown due to a lack of correlation between YLF and disease severity (Sarovich et al., 2014). Animal studies to determine the virulence of these and other African isolates would be a worthwhile endeavor in follow-up to this work.

The two-step selective enrichment process described by Trinh and colleagues (Trinh et al., 2019) proved to be a robust and reproducible method to culture *B. pseudomallei* from these complex environmental samples. This method employed the traditional consensus guidelines but also capitalized on *B. pseudomallei's* unique ability to utilize erythritol as its sole carbon source. While near neighbor species were more-than-likely present in the original soil sample, every colony we examined that morphologically resembled *B. pseudomallei* on Ashdown's agar was positively confirmed to be *B. pseudomallei* by PCR. We recommend this selective enrichment method be adapted more widely for future environmental surveys.

The development of sustainable rice farming is expanding in Ghana. In the Ashanti Region, the southern part of Ghana, more than one hundred hectares of land have been manually developed for rice crops since 2004. Considerable development has occurred to date and the Inland Valley Rice Development Project (IVRDP) aims to expand land development by another 1,500 ha, including Ejisu-Juaben, Ahafo Ano South, Ahafo Ano North, and Ejura-Sekyedumasi districts of the Ashanti region. These developments will lead to land use changes including clearing of sites, leveling, and terracing of rice fields, development of water control structures and development of access tracks (Ministry of Food and Agriculture, 2019). The IVRDP Rice development project has led to significant positive effects on communities, including income generation and education. Changes in land use may also have contributed to changes seen in human disease prevalence such as malaria, a disease that is readily recognized by routine laboratory testing (Mpianing, 2016).

However, undifferentiated clinical diseases not subjected to specific laboratory testing may remain uncharacterized, as evidenced by a recent report out of a hospital in the region that details 200 gram-negative clinical isolates obtained over a six-month period that does not include *B. pseudomallei* (Agyepong et al., 2018).

The emergence of new pathogens or unmasking of previously underappreciated pathogens provides the opportunity to increase the capacity to diagnose and treat the infections they cause (Birnie et al., 2022). We performed phenotypic analyses that demonstrate that these novel isolates conform to common characteristics such as CPS production detectible by the Active Melioidosis Detect diagnostic, and one of the most common *B. pseudomallei* metabolic profiles differentiated by API 20NE (Table 3) (Amornchai et al., 2007; Houghton et al., 2014). Each isolate was also susceptible to common antibiotics used in the management of melioidosis (Table 4).

In conclusion, environmental testing plays an important role in defining geographical regions with increased risk of melioidosis, as well as to provide unique genomes for in-depth phylogenetic analyses and epidemiological investigations. Through soil sampling and genomic analyses, we identified a novel strain type of *B. pseudomallei* in Africa that contains a putative virulence factor typically associated with clinical isolates in Thailand. Environmental samples in this region represent a rich, but as-yet-untapped source for future endemicity studies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found here: https://www.ncbi.nlm.nih.gov/, PRJNA1078842.

Author contributions

KS: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. LV: Data curation, Formal analysis, Methodology, Writing – review & editing. GR: Data curation, Formal analysis, Investigation, Software, Writing – review & editing. HD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing. MF: Data curation, Formal analysis, Investigation, Software, Writing – review & editing. FM: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Writing – review & editing. AS: Data curation, Formal analysis, Investigation, Writing – review & editing. MD: Data curation, Formal analysis, Investigation, Writing – review & editing. GO: Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. FR: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing. PD: Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. AO-O: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. DC: Investigation, Resources, Supervision, Writing – review & editing. RC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. KB-L: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. Field work was conducted under the Joint West Africa Research Group funded through the Defense Health Program with programmatic oversight from the Military Infectious Diseases Research Program. Funding for the genomic characterization was provided through the Armed Forces Health Surveillance Division (AFHSD), Global Emerging Infections Surveillance (GEIS) Branch, ProMIS ID P0008_22_NM.

Acknowledgments

We thank the Chief and Elders of Nobewam for their approval and community access, and the Rice Farmers of Nobewam for access to their land. We greatly appreciate the Komfo Anokye Teaching Hospital (KATH) staff for their participation and logistical support of this study. We especially thank Chris Oppong, Thomas K. Gyampomah, Isaac Boakye, and Emmanuel Quainoo from KATH.

Author disclaimer

KB-L, RC, and KS were federal employees of the United States government. This work was prepared as part of their official duties. Title 17 U.S.C. 105 provides that "copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. 101 defines a US Government work as work prepared by a military service member or employee of the US Government as part of that person's official duties. The views expressed in this work reflect the results of research conducted by the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

Conflict of interest

LV, GR, HD, MF, and FM were employed by Leidos.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1401259/full#supplementary-material

References

Adhikari, B. N., Paskey, A. C., Frey, K. G., Bennett, A. J., Long, K. A., Kuhn, J. H., et al. (2024). Virome profiling of fig wasps (*Ceratosolen* spp.) reveals virus diversity spanning four realms. *Virology* 591:109992. doi: 10.1016/j.virol.2024.109992

Agyepong, N., Govinden, U., Owusu-Ofori, A., and Essack, S. Y. (2018). Multidrug-resistant gram-negative bacterial infections in a teaching hospital in Ghana. *Antimicrob. Resist. Infect. Control* 7:37. doi: 10.1186/s13756-018-0324-2

Amornchai, P., Chierakul, W., Wuthiekanun, V., Mahakhunkijcharoen, Y., Phetsouvanh, R., Currie, B. J., et al. (2007). Accuracy of *Burkholderia pseudomallei* identification using the API 20NE system and a latex agglutination test. *J. Clin. Microbiol.* 45, 3774–3776. doi: 10.1128/JCM.00935-07

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.

Birnie, E., James, A., Peters, F., Olajumoke, M., Traore, T., Bertherat, E., et al. (2022). Melioidosis in Africa: Time to raise awareness and build capacity for its detection, diagnosis, and treatment. *Am. J. Trop. Med. Hyg.* 106, 394–397. doi: 10.4269/ajtmh.21-0673

Birnie, E., Van 't Hof, S., Bijnsdorp, A., Mansaray, Y., Huizenga, E., Van Der Ende, A., et al. (2019). Identification of Burkholderia thailandensis with novel genotypes in the soil of central Sierra Leone. *PLoS Negl. Trop. Dis.* 13:e0007402. doi: 10.1371/journal.pntd.0007402

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.

Bushnell, B. (2014). BBMAP: A fast, accurate, splice-aware aligner, USDOE Joint Genome Institute (JGI): Walnut Creek.

Chewapreecha, C., Holden, M. T., Vehkala, M., Valimaki, N., Yang, Z., Harris, S. R., et al. (2017). Global and regional dissemination and evolution of Burkholderia pseudomallei. *Nat. Microbiol.* 2:16263.

Cuadros, J., Gil, H., Miguel, J. D., Marabe, G., Gomez-Herruz, T. A., Lobo, B., et al. (2011). Case report: Melioidosis imported from West Africa to Europe. *Am. J. Trop. Med. Hyg.* 85, 282–284. doi: 10.4269/ajtmh.2011.11-0207

Dale, J., Price, E. P., Hornstra, H., Busch, J. D., Mayo, M., Godoy, D., et al. (2011). Epidemiological tracking and population assignment of the non-clonal bacterium, Burkholderia pseudomallei. *PLoS Negl. Trop. Dis.* 5:e1381. doi: 10.1371/journal.pntd.0001381

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403.

Gee, J. E., Gulvik, C. A., Elrod, M. G., Batra, D., Rowe, L. A., Sheth, M., et al. (2017). Phylogeography of Burkholderia pseudomallei Isolates, Western Hemisphere. *Emerg. Infect. Dis.* 23, 1133–1138. doi: 10.3201/eid2307.161978

Godoy, D., Randle, G., Simpson, A. J., Aanensen, D. M., Pitt, T. L., Kinoshita, R., et al. (2003). Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J. Clin. Microbiol.* 41, 2068–2079. doi: 10.1128/JCM.41.5.2068-2079.2003

GPS Coordinates,. (2023). *Distance between coordinates: Distance calculator*. Available online at: https://gps-coordinates.org/distance-between-coordinates.php (accessed January 12, 2024).

Hall, C. M., Jaramillo, S., Jimenez, R., Stone, N. E., Centner, H., Busch, J. D., et al. (2019). *Burkholderia pseudomallei*, the causative agent of melioidosis, is rare but ecologically established and widely dispersed in the environment in Puerto Rico. *PLoS Negl. Trop. Dis.* 13:e0007727. doi: 10.1371/journal.pntd.0007727

Houghton, R. L., Reed, D. E., Hubbard, M. A., Dillon, M. J., Chen, H., Currie, B. J., et al. (2014). Development of a prototype lateral flow immunoassay (LFI) for the rapid diagnosis of melioidosis. *PLoS Negl. Trop. Dis.* 8:e2727. doi: 10.1371/journal.pntd.0002727

Jayasinghearachchi, H. S., Muthugama, T. A., Masakorala, J., Kulasekara, U. S., Jayaratne, K., Jayatunga, D., et al. (2022). *Burkholderia pseudomallei* in soil and natural water bodies in rural Sri Lanka: A hidden threat to public health. *Front. Vet. Sci.* 9:1045088. doi: 10.3389/fvets.2022.1045088

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., and Huttenhower, C. (2015). High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* 11:e1004557. doi: 10.1371/journal.pcbi.1004557

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546.

Li, P. E., Lo, C. C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* 45, 67–80. doi: 10.1093/nar/gkw1027

Limmathurotsakul, D., Golding, N., Dance, D. A., Messina, J. P., Pigott, D. M., Moyes, C. L., et al. (2016). Predicted global distribution of Burkholderia pseudomallei and burden of melioidosis. *Nat. Microbiol.* 1:15008.

Mabayoje, D. A., Kenna, D. T. D., Dance, D. A. B., and Nicfhogartaigh, C. (2022). Melioidosis manifesting as chronic femoral osteomyelitis in patient from Ghana. *Emerg. Infect. Dis.* 28, 201–204. doi: 10.3201/eid2801.211800

Ministry of Food and Agriculture (2019). *Inland valley rice development project (Republic of Ghana)*. Available online at: https://mofa.gov.gh/site/projects/49-inland-valley-rice-dissemination-project-ndrp (Accessed January 17 2024).

Morosini, M. I., Quereda, C., Gil, H., Anda, P., Nunez-Murga, M., Canton, R., et al. (2013). Melioidosis in traveler from Africa to Spain. *Emerg. Infect. Dis.* 19, 1656–1659.

Mpianing, E. (2016). Effects of inland valley rice development project on household poverty indicators of rice farmers in ahafo-ANO south district in the ashanti region Of Ghana. *Int. J. Sci. Technol. Res.* 5, 14–23.

Mukhopadhyay, C., Shaw, T., Varghese, G. M., and Dance, D. A. B. (2018). Melioidosis in South Asia (India, Nepal, Pakistan, Bhutan and Afghanistan). *Trop. Med. Infect. Dis.* 3:51. doi: 10.3390/tropicalmed3020051

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Novak, R. T., Glass, M. B., Gee, J. E., Gal, D., Mayo, M. J., Currie, B. J., et al. (2006). Development and evaluation of a real-time PCR assay targeting the type III secretion system of *Burkholderia pseudomallei*. *J. Clin. Microbiol.* 44, 85–90. doi: 10.1128/JCM.44.1.85-90.2006

Oduro, G., Robberts, F. J. L., Dartey, P. K. A., Owusu-Ofori, A., Oppong, C., Gyampomah, T. K., et al. (2022). On the environmental presence of *Burkholderia pseudomallei* in South-Central Ghana. *Appl. Environ. Microbiol.* 88:e0060022. doi: 10.1128/aem.00600-22

Roe, C., Vazquez, A. J., Phillips, P. D., Allender, C. J., Bowen, R. A., Nottingham, R. D., et al. (2022). Multiple phylogenetically-diverse, differentially-virulent *Burkholderia pseudomallei* isolated from a single soil sample collected in Thailand. *PLoS Negl. Trop. Dis.* 16:e0010172. doi: 10.1371/journal.pntd.0010172

Salam, A. P., Khan, N., Malnick, H., Kenna, D. T., Dance, D. A., and Klein, J. L. (2011). Melioidosis acquired by traveler to Nigeria. *Emerg. Infect. Dis.* 17, 1296–1298. doi: 10.3201/eid1707.100502

Sarovich, D. S., Garin, B., De Smet, B., Kaestli, M., Mayo, M., Vandamme, P., et al. (2016). Phylogenomic analysis reveals an asian origin for African *Burkholderia pseudomallei* and further supports melioidosis endemicity in Africa. *mSphere* 1:e00089-15. doi: 10.1128/mSphere.00089-15

Sarovich, D. S., Price, E. P., Webb, J. R., Ward, L. M., Voutsinos, M. Y., Tuanyok, A., et al. (2014). Variable virulence factors in *Burkholderia pseudomallei* (melioidosis) associated with human disease. *PLoS One* 9:e91682. doi: 10.1371/journal.pone.0091682

Savelkoel, J., Oladele, R. O., Ojide, C. K., Peters, R. F., Notermans, D. W., Makinwa, J. O., et al. (2023). Presence of Burkholderia pseudomallei in Soil, Nigeria, 2019. *Emerg. Infect. Dis.* 29, 1073–1075. doi: 10.3201/eid2905.221138

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069.

Seemann, T. (2015). *Snippy: Rapid haploid variant calling and core SNP phylogeny*, 4th Edn. Available online at: https://github.com/tseemann/snippy (accessed January 12, 2024).

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res* 34, D32–D36.

Steinmetz, I., Wagner, G. E., Kanyala, E., Sawadogo, M., Soumeya, H., Teferi, M., et al. (2018). Melioidosis in Africa: Time to uncover the true disease load. *Trop Med Infect Dis* 3:62. doi: 10.3390/tropicalmed3020062

Stone, N. E., Hall, C. M., Browne, A. S., Sahl, J. W., Hutton, S. M., Santana-Propper, E., et al. (2020). *Burkholderia pseudomallei* in Soil, US Virgin Islands, 2019. *Emerg. Infect. Dis.* 26, 2773–2775. doi: 10.3201/eid2611.191577

Torres, A. G. (2023). The public health significance of finding autochthonous melioidosis cases in the continental United States. *PLoS Negl. Trop. Dis.* 17:e0011550. doi: 10.1371/journal.pntd.0011550

Trinh, T. T., Assig, K., Tran, Q. T. L., Goehler, A., Bui, L. N. H., Wiede, C., et al. (2019). Erythritol as a single carbon source improves cultural isolation of *Burkholderia pseudomallei* from rice paddy soils. *PLoS Negl. Trop. Dis.* 13:e0007821. doi: 10.1371/journal.pntd.0007821

Tuanyok, A., Leadem, B. R., Auerbach, R. K., Beckstrom-Sternberg, S. M., Beckstrom-Sternberg, J. S., Mayo, M., et al. (2008). Genomic islands from five strains of *Burkholderia pseudomallei*. *BMC Genom.* 9:566. doi: 10.1186/1471-2164-9-566

Wall, R. A., Mabey, D. C., Corrah, P. T., and Peters, L. (1985). A case of melioidosis in West Africa. *J. Infect. Dis.* 152, 424–425.

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

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 July 2024
DOI 10.3389/fmicb.2024.1437572

OPEN ACCESS

EDITED BY
Gabriel Luz Wallau,
Aggeu Magalhães Institute (IAM), Brazil

REVIEWED BY
Lara Ferrero Gómez,
Universidade Jean Piaget de Cabo,
Cabo Verde
Mary Bosch Farone,
Middle Tennessee State University,
United States

*CORRESPONDENCE
Robert P. Hirt
✉ Robert.Hirt@ncl.ac.uk

RECEIVED 23 May 2024
ACCEPTED 28 June 2024
PUBLISHED 17 July 2024

CITATION
Mpeyako LA, Hart AJ, Bailey NP, Carlton JM, Henrissat B, Sullivan SA and Hirt RP (2024) Comparative genomics between *Trichomonas tenax* and *Trichomonas vaginalis*: CAZymes and candidate virulence factors.
Front. Microbiol. 15:1437572.
doi: 10.3389/fmicb.2024.1437572

Comparative genomics between *Trichomonas tenax* and *Trichomonas vaginalis*: CAZymes and candidate virulence factors

Lenshina A. Mpeyako[1], Adam J. Hart[1], Nicholas P. Bailey[1], Jane M. Carlton[2,3], Bernard Henrissat[4,5], Steven A. Sullivan[2,3] and Robert P. Hirt[1]*

[1]Biosciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom, [2]Department of Biology, Center for Genomics and Systems Biology, New York University, New York, NY, United States, [3]Department of Molecular Microbiology and Immunology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States, [4]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, [5]Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark, Lyngby, Denmark

Introduction: The oral trichomonad *Trichomonas tenax* is increasingly appreciated as a likely contributor to periodontitis, a chronic inflammatory disease induced by dysbiotic microbiota, in humans and domestic animals and is strongly associated with its worst prognosis. Our current understanding of the molecular basis of *T. tenax* interactions with host cells and the microbiota of the oral cavity are still rather limited. One laboratory strain of T. tenax (Hs-4:NIH/ATCC 30207) can be grown axenically and two draft genome assemblies have been published for that strain, although the structural and functional annotation of these genomes is not available.

Methods: GenSAS and Galaxy were used to annotate two publicly available draft genomes for *T. tenax*, with a focus on protein-coding genes. A custom pipeline was used to annotate the CAZymes for *T. tenax* and the human sexually transmitted parasite *Trichomonas vaginalis*, the most well-characterized trichomonad. A combination of bioinformatics analyses was used to screen for homologs of *T. vaginalis* virulence and colonization factors within the *T. tenax* annotated proteins.

Results: Our annotation of the two *T. tenax* draft genome sequences and their comparison with *T. vaginalis* proteins provide evidence for several candidate virulence factors. These include candidate surface proteins, secreted proteins and enzymes mediating potential interactions with host cells and/or members of the oral microbiota. The CAZymes annotation identified a broad range of glycoside hydrolase (GH) families, with the majority of these being shared between the two *Trichomonas* species.

Discussion: The presence of candidate *T. tenax* virulence genes supports the hypothesis that this species is associated with periodontitis through direct and indirect mechanisms. Notably, several GH proteins could represent potential new virulence factors for both Trichomonas species. These data support a model where *T. tenax* interactions with host cells and members of the oral microbiota could synergistically contribute to the damaging inflammation characteristic of periodontitis, supporting a causal link between *T. tenax* and periodontitis.

KEYWORDS

protein-coding genes, surface proteins, secreted proteins, glycoside/glycosyl hydrolases (GHs), peptidases/proteases, exosomes, pore-forming proteins

1 Introduction

The trichomonad *Trichomonas tenax* (Phylum Parabasalia, Class Trichomonadea, Cepicka et al., 2010) is an obligate symbiont of the oral cavity reported to have a potential role in periodontitis (Marty et al., 2017; Santi-Rocca, 2020; Eslahi et al., 2021; Martin-Garcia et al., 2022). However, *T. tenax* is one of the least well-understood microbial species of the complex oral microbial ecosystem and very little is known on how it might modulate human oral disease (Baker et al., 2024). Similar to the closely related sexually transmitted parasite *Trichomonas vaginalis*, which is human-specific, *T. tenax* was also commonly considered to be specifically associated with humans (Honigberg, 1990; Maritz et al., 2014). However, the use of sensitive molecular diagnostic tools has clearly demonstrated that *T. tenax* is common in the oral cavity of a broader range of hosts including, cats, dogs and horses and with some additional reports among birds (Kellerová and Tachezy, 2017; Matthew et al., 2023). There are also reports of the isolation of *T. tenax* from extra-oral locations including the respiratory tract, lymph nodes, salivary glands and the urogenital tract (Maritz et al., 2014; Brosh-Nissimov et al., 2019). Studies have reported an increased prevalence of *T. tenax* among patients with periodontitis compared to healthy controls (Marty et al., 2017; Eslahi et al., 2021; Martin-Garcia et al., 2022). However, the ability of *T. tenax* to contribute, directly or indirectly, to the pathobiology of periodontitis, is still poorly understood. Indeed, there is a lack of data on the molecular and cellular basis of *T. tenax* interactions with host cells (Matthew et al., 2023), members of the microbiota, and its potential endosymbionts, three key factors considered important in modulating the pathobiology of *T. vaginalis* (Hirt et al., 2011; Mercer and Johnson, 2018; Riestra et al., 2022; Margarita et al., 2023).

Trichomonad genomes typically exceed 50 Mb, making them unusually large compared to other parasitic microbial eukaryotes (Zubáčová et al., 2008). The published draft genome data for *T. vaginalis* (G3) is consistent with these estimates, with a size of ~160 Mb (Carlton et al., 2007). The first draft whole genome sequence for *T. tenax* (strain Hs-4:NIH, ATCC 30207; Manassas, VA, USA) was published in 2019 (~47 Mb, across 4,161 contigs) (Benabdelkader et al., 2019). Benabdelkader et al. (2019) conducted this study to investigate the prevalence and genetic diversity of *T. tenax* and its potential involvement in the severity of periodontitis among humans. A second draft whole genome sequence dataset for the same strain of *T. tenax* was published in 2022 (63.4 Mb, across 4906 contigs) (Yang et al., 2022). Both draft genomes are smaller than the 133 ± 4 Mb estimated from flow cytometry (Zubáčová et al., 2008). The two published *T. tenax* draft genomes did not include annotation of protein-coding genes, which is essential to guide the study of the molecular cell biology of a given organism. Here, we investigated *T. tenax* protein-coding genes by integrating and annotating the two published draft genomes (Benabdelkader et al., 2019; Yang et al., 2022). The predicted *T. tenax* protein-coding genes and their functional annotations were compared with the predicted proteome of a new assembly of *T. vaginalis*. We also annotated the CAZymes (Lombard et al., 2014; Wardman et al., 2022) from *T. tenax* and *T. vaginalis*. We focused our analyses to specifically test the hypothesis that *T. tenax* encodes a range of virulence and colonization factors that contribute directly and indirectly to the pathobiology of periodontitis through interaction with various host cells and members of the oral microbiota, as known for *T. vaginalis* (Hirt, 2013; Mercer and Johnson, 2018; Riestra et al., 2022). Indeed, only a handful of publications have investigated the potential virulence of *T. tenax* on host cells and the virulence factors that could be associated with the observed impact on host cells (reviewed in Matthew et al., 2023). Similarly, we are aware of only one study that has considered *T. tenax* genes encoding candidate proteins mediating interactions with members of the microbiota, i.e. the *T. vaginalis* NlpC/P60 peptidases that target bacterial cell walls and are conserved in *Trichomonas gallinae* (Barnett et al., 2023). Here we took advantage of the rich annotation available for *T. vaginalis* virulence factors (Carlton et al., 2007; Hirt et al., 2011; Hirt, 2013; Mercer and Johnson, 2018; Riestra et al., 2022) to expand the bioinformatic characterization of *T. tenax* candidate virulence factors. These could mediate binding of *T. tenax* to human epithelial cells, immunocytes, the microbiota in the oral cavity, or could degrade host structural proteins or those involved in innate and adaptive immunity, and by doing so could contribute to the pathobiology of periodontitis.

2 Materials and methods

2.1 Genome annotation

GenSAS (Humann et al., 2019) was used to annotate the two draft genome sequences of *T. tenax* strain Hs-4:NIH (NCBI Genome assemblies PRJEB22701 and ASM2309173v1) (Benabdelkader et al., 2019; Yang et al., 2022). Identification and masking of repeats was achieved using RepeatMasker (Nishimura, 2000) and RepeatModeler (Smit and Hubley, 2008) on GenSAS. Multiple tools integrated into GenSAS were employed for the *ab initio* prediction of protein-coding genes including Augustus (Stanke et al., 2004), GeneMarkES (Lomsadze et al., 2005), GlimmerM (Salzberg et al., 1999), Genescan (Burge and Karlin, 1997) and SNAP (Korf, 2004). For homology-based gene predictions protein and transcript sequences from other trichomonads, including *T. vaginalis* (Carlton et al., 2007), and other relevant eukaryotic species were downloaded from NCBI and aligned to the genomes using BLAST (Altschul et al., 1990), BLAT (Kent, 2002), PASA (Haas et al., 2003) and DIAMOND (Buchfink et al., 2021) using default parameters. A published *T. tenax* RNA–Seq dataset (Handrich et al., 2019) (NCBI SRA accession SRX2052871) was used as evidence for expression of protein-coding genes. The RNA–Seq reads were aligned to the repeat masked genome assembly using HISAT2 (Kim et al., 2015) and TopHat2 (Kim et al., 2013) with default parameters. EvidenceModeler (Haas et al., 2008) was used to create the consensus gene set by incorporating the outputs of all *ab initio* and alignment-based gene predictions. The official gene set (OGS) which is defined as the set of predictions that serve as the most trusted gene predictions, was generated by PASA Refinement (Haas et al., 2003). In this study, all outputs generated from the gene prediction tools were included and integrated to generate a consensus annotation.

2.2 Protein functional annotation

For protein functional annotation, tools integrated into GenSAS were used and included DIAMOND (Buchfink et al., 2021) and InterProScan (Jones et al., 2014). These analyses were complemented with tools integrated into Galaxy (Afgan et al., 2016) and included BlastP (Camacho et al., 2009), EggNog Mapper (Huerta-Cepas et al., 2016) and Motif Search (Kanehisa, 2002), PANZZER2 (Törönen et al., 2018) and GhostKoala (Kanehisa et al., 2016). The predicted proteins were used to query the NCBI non-redundant (nr) refseq_protozoa protein database (released on 9/12/2021) using the BLASTP search, with an E-value cut-off of 1×10^{-8}. That version of the database contains 1,095,419 protein sequences from different protozoans including *T. vaginalis*. The inferred protein sequences of *T. tenax* were also used to query the SWISS-PROT and TrEMBL databases using DIAMOND to further improve the accuracy of function allocation. Gene name assignment was done by recording BLASTP top hits and PANZZER2 (Törönen et al., 2018). Additional functional domains were predicted by EggNog Mapper (Huerta-Cepas et al., 2017). Default settings were used with a minimum E-value expected to be 1×10^{-3}. Both InterProScan and Eggnog Mapper provided additional evidence for existing gene annotations including the database cross-references (Dbxref) and gene ontology terms (GO). We also used SMART (Letunic and Bork, 2018) to investigate the domain organization of specific proteins of interest. To further enrich the protein functional annotation, GhostKOALA (Kanehisa et al., 2016) was used to assign K numbers to the annotations by GHOSTX search against a nonredundant set of KEGG GENES. The GhostKOALA platform is a web-based server that performs automatic annotation of genomes with Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (Kanehisa et al., 2016). Protein localization was inferred using Phobius (Käll et al., 2007), SignalP (Petersen et al., 2011) and for selected entries DeepTMHMM (Hallgren et al., 2022). Data produced by the different software were mined and an annotation summary generated manually.

In addition to the published 2007 annotation of the *T. vaginalis* G3 draft genome (Carlton et al., 2007), we leveraged updated annotation from a new *T. vaginalis* G3 assembly featuring six chromosome-length scaffolds derived from long-read sequencing. The new genome sequence data and corresponding assembly and annotation are all available at the NCBI [NCBI: Name, NYU_TvagG3_2, BioProject PRJNA16084, NCBI RefSeq assembly GCF_026262505.1; also TrichDB (Aurrecoechea et al., 2009): release 65, 14/09/2023]. These updated data are referred to here as the G3_2022 data. Supplementary Table 1 provides a lookup table between the G3_2007 and G3_2022 annotations providing continuity with the previous assembly by including where possible the TrichDB IDs (i.e., TVAG_XXXXXX, TVAG_RG_XXXXX).

2.3 CAZy annotation

The general protein functional annotation from *T. tenax* (this study) and *T. vaginalis* (Carlton et al., 2007) genomes were enriched by submitting the predicted protein sequences to the CAZy annotation pipeline. The reported data in this study have a specific focus on glycoside hydrolases (GH), key Carbohydrate Active enZymes (CAZymes), and the carbohydrate-binding modules (CBMs), as these are more likely to represent potential virulence factors. Briefly, automated crunching was performed followed by manual curation of borderline cases and fragments as described for hundreds of annotated genomes and used for updating the CAZy database (Lombard et al., 2014). Predictions of enzyme activities for *Trichomonas* spp. GHs, based on significant similarities with entries from a library of experimentally determined enzymatic activities of CAZymes (Drula et al., 2022), are also reported (Drula et al., 2022).

2.4 Estimates of completeness

Genome annotation quality and the completeness of gene annotation were estimated and quantified using the Benchmarking for Universal Single-Copy Orthologs (BUSCO) tool (Manni et al., 2021). The annotated proteins of *T. tenax* were searched against selected near-universal single-copy orthologs from the eukaryotic database (eukaryota_odb10 v5.2.2) of the hierarchical catalog of orthologs (OrthoDB v10) (Kriventseva et al., 2007). BUSCO searches the query proteome for "complete – single copy," "complete – duplicated," "fragmented," and "missing" orthologs within the proteome that are expected to be highly conserved among related species. The output generated from this evaluation is presented as the percentage of the total 255 BUSCOs in the eukaryote_odb10 lineage from the OrthoDB database.

2.5 Gene family analyses

We employed OrthoVenn3 (using the Orthofinder algorithm) (Emms and Kelly, 2015; Sun et al., 2023) to identify and compare orthologs in the protein sequence sets from *T. tenax* (integrated annotated proteins) and *T. vaginalis*. Orthofinder utilizes DIAMOND for identifying sequence similarity and DendroBLAST (Kelly and Maini, 2013) for gene tree inference.

We also performed a gene network analysis using EGN (Halary et al., 2013) to predict homologous gene families using the full list of annotated proteins from the genomes of *T. vaginalis* G3_2022 (Genbank accession GCA_026262505.1) and *T. tenax* (this study). Results of an all vs all BLASTP searches (E-value $\leq 1 \times 10^{-10}$, percent identity >25%, aligned length of query and subject $\geq$90%) were clustered into networks with edges linking gene nodes according to BLAST hits. Thresholds for recording significant BLAST hits were an E-value of less than 1×10^{-10}, percentage identity of greater than 25% and an alignment length of at least 90% of the length of both sequences. A selection of gene networks for complex gene families of interest was visualized. Figures were generated using Cytoscape using the yfiles organic layout algorithm (Shannon et al., 2003).

TABLE 1 Overview of selected characteristics and annotations for the two published *Trichomonas tenax* draft genome sequences.

Draft genome*	*T. tenax* (2019)	*T. tenax* (2022)
Isolate/Strain	Hs-4:NIH	
Assembly size (Mb)	46.74	63.38
No. of scaffolds	4,161	4,904
N50 (bp)	13,554	38,386
L50 (bp)	1,070	469
% GC (Genome)	34.6	33.70
Genome assemblies ID	PRJEB22701	ASM2309173v1
No. of protein coding genes	16,786	18,852
BUSCO score (complete)**	129	
BUSCO score (Missing or Fragmented)**	126	
BUSCO score (Complete %)**	50.6	
tRNAs	226	298
rRNAs	2	4

*Benabdelkader et al. (2019) and Yang et al. (2022).

**BUSCO score based on the integrated annotations (20,287 proteins) from the 2019 and 2022 draft genomes.

3 Results

3.1 Gene predictions, functional annotation and completeness of *T. tenax* annotations

An overview of the annotation of the two *T. tenax* Hs-4:NIH draft genome sequences and their level of completeness is provided in Table 1. The Supplementary Data 1, 2 are the GFF3 files of the annotation of the *T. tenax* 2019 and 2022 draft genomes, respectively. Integrating the two annotations predicted a total of 20,286 distinct protein-coding genes for the *T. tenax* strain Hs-4:NIH. A FASTA file with the 20,286 protein sequences is available as the Supplementary Data 3. The BUSCO completeness score of 50.6% for *T. tenax* integrated annotation (20,287 proteins) is similar to the *T. vaginalis* (2022 annotation), with a completeness score of 53%. Here we describe the functional annotation of a selection of *T. tenax* proteins with a focus on CAZymes and a selection of some of the best characterized candidate virulence factors by taking advantage of the more extensive annotation of *T. vaginalis* proteins (Carlton et al., 2007, 2010; Hirt et al., 2011; Riestra et al., 2022). Supplementary Table 2 lists the 20,286 *T. tenax* (strain Hs-4:NIH) proteins with their functional annotations and inferred protein sequences, derived from the integrated annotation of the 2019 and 2022 draft genome sequence data.

3.2 CAZy annotation for *T. tenax* and *T. vaginalis*

A specific CAZy annotation was performed for both *T. tenax* (annotation from this study) and *T. vaginalis* (G3_2007 annotation) (Supplementary Table 3). Only 21 *T. vaginalis* CAZy annotated entries (7%), for a total of 287 entries, were deprecated, or not yet vetted, in the most recent assembly and corresponding annotation, G3_2022 annotation (Supplementary Tables 1, 3). The majority of *Trichomonas* spp. CAZy entries correspond to CAZymes members of the GH class with 126 TtGHs and 181 TvGHs. There are 86 *T. tenax* glycosyltransferases (TtGT) and 91 *T. vaginalis* glycosyltransferases (TvGT) (Supplementary Table 3). There are far fewer polysaccharide lyases (PLs) with three TtPL and 14 TvPL (Supplementary Table 3). Three *T. tenax* proteins are predicted to contain a CBM domain only, without an identified CAZyme domain (two entries with one CBM20, and one entry with one CBM32). Similarly, there are only three proteins with a single CBM13 domain without an identified CAZyme domain in *T. vaginalis* (Supplementary Table 3). Of the 189 GH families currently recognized in CAZy (http://www.cazy.org/ April, 25th 2024) (named GH1 to GH189), a total of 26 families were identified among either of the two species (Table 2, Supplementary Table 3). Some of the *Trichomonas* spp. GH proteins have evidence of a signal peptide (SP) and/or a transmembrane domain(s) (TMD), suggesting that these CAZymes could act on glycans with extracellular locations or within an organelle (Table 2). Among the GH13 proteins (14 in *T. tenax* and 24 in *T. vaginalis*), two in each species contain a GH133 domain (Table 2, Supplementary Table 3). Seven families with 10 or more, entries in at least one *Trichomonas* species include the GH13, GH16, GH30, GH31, GH47, GH77 and GH163 (Table 2). Table 3 lists selected features for the GH families with the largest number of entries or that include one or more activities that comprise potential mucin-targeting enzymes (mucinases) (Labourel et al., 2023) or other host glycoproteins encoded by both *Trichomonas* species. Two families, GH5 and GH32, are specific for *T. vaginalis* with one entry for each family, and these are confirmed in the G3_2022 annotation (Table 2, Supplementary Table 3). Table 4 lists selected characteristics for these two *T. vaginalis* specific GHs families.

3.3 Identification of candidate virulence factors in *T. tenax* through comparisons with *T. vaginalis*

3.3.1 Peptidases

Similar to the 469 predicted peptidases in *T. vaginalis* (Carlton et al., 2007), a large number of candidate peptidases (476) were identified in *T. tenax* (Supplementary Table 4). These include 240 cysteine peptidases, 141 metallopeptidases, 85 serine peptidases and seven aspartic peptidases (Supplementary Table 4). A selection of *T. tenax* peptidase families, including some of those with the largest number of homologs to *T. vaginalis* factors and that are experimentally demonstrated to contribute to host damage or are candidate virulence factors, are listed in Table 5. A number of these peptidases represent candidate secreted or cell surface proteins based on the presence of inferred signal peptides and/or TMD, consistent with a potential role in targeting extracellular or phagocytosed proteins from host or microbiota cells (including peptidoglycans) (Supplementary Table 4, Table 5). The *T. tenax* genome encodes a large number (100 members) of family C19

TABLE 2 Annotated candidate glycoside hydrolases (GH) for *T. tenax and T. vaginalis*.

GH family #*	*Trichomonas tenax* (Hs-4:NIH)				*Trichomonas vaginalis* (G3_2007)			
	Number annotated	SP	TMD	SP + TMD	Number annotated	SP	TMD	SP + TMD
GH1	2	1	1	0	2	0	1	0
GH2	6	2	0	0	5	2	0	0
GH3	7	2	1	0	4	0	1	0
GH5[a]	0	0	0	0	1	0	0	0
GH13	14	4	4	0	23	5	4	0
GH14	4	3	0	0	4	2	0	0
GH15	1	1	0	0	3	1	0	0
GH16	2	1	0	1	10	0	0	1
GH18	1	0	0	0	1	0	0	0
GH19	6	2	1	0	8	0	0	0
GH20	6	2	2	1	6	0	2	0
GH25	5	0	0	0	4	0	0	0
GH27	2	1	1	0	1	0	0	0
GH29	2	1	0	1	2	0	2	0
GH30	1	1	0	0	16	0	0	0
GH31	17	6	2	3	17	3	2	1
GH32[a]	0	0	0	0	1	0	0	0
GH33	1	1	0	0	2	0	0	0
GH36	2	1	1	0	2	0	1	0
GH37	1	0	0	0	2	0	0	0
GH38	4	0	2	0	5	0	0	0
GH47	24	11	8	1	27	1	14	0
GH77	7	0	1	0	14	0	0	0
GH99	8	3	1	0	6	1	2	0
GH133 (GH13)	2	0	0	0	2	0	0	0
GH163	2	0	1	1	10	1	4	3
Total GH domains**	127	43	26	8	178	16	33	5

*SP, signal peptide; TMD, transmembrane domain.

[a] GH families restricted to *T. vaginalis*.

**Number of GH families corresponding to 125 and 176 distinct proteins for *T. tenax* and *T. vaginalis*, respectively, with two proteins with the configuration GH13-GH133 in each species.

ubiquitin hydrolases, similar to *T. vaginalis* (117 members) (Carlton et al., 2007) (Supplementary Table 4). In *T. vaginalis* > 50% of the C19 ubiquitin hydrolases are likely to be functional (Hirt et al., 2011). This large number of C19 peptidases that are known to be functionally associated with the 20S proteasome (ubiquitin C-terminal hydrolases) in model systems (Barrett and Rawlings, 2001) highlight for these two species the importance of cytosolic protein degradation.

3.3.2 Surface proteins as adhesins

A large number of genes were annotated *in T. vaginalis* to encode candidate surface proteins mediating key interactions with its environment, including host epithelial cells (adhesins) and members of the microbiota (Carlton et al., 2007; Hirt et al., 2011; Riestra et al., 2022). Several of these proteins have been experimentally demonstrated in *T. vaginalis* to be expressed on the cell surface and to mediate binding to host cells (Hirt et al., 2011; Riestra et al., 2022). We focus here on some specific examples which indicate that *T. tenax* has a similar repertoire of candidate binding factors to host cells, and potentially to members of the oral microbiota. These include 244 TtBspA-like (Supplementary Table 5) and 24 TtPmp-like (Supplementary Table 6) proteins (Figure 1). Two TtBspA-like proteins contain a peptidase domain, including an M8 peptidase and a NlpC/P60 peptidase (Figure 2).

TABLE 3 Selected characteristics of glycoside hydrolase families with 10 or more entries or known to include enzymes targeting mucins or other host glycoproteins, encoded by both *T. tenax and T. vaginalis*.

GH family #*	Known enzymatic activities(s)	Comments	References
GH2 (6/5)	16 different known activities, including β-galactosidases and β -mannosidase, the two predicted activities for the enzymes of the two species	Mostly bacterial and a few fungal enzymes. Include mucinases.	Lombard et al., 2014; Labourel et al., 2023
GH16 (2/10)	21 different activities, including endo-β-1,4-galactosidases and endo-β-1,3-glucosidases	Mostly bacterial and some fungal enzymes. Classified in 23 subfamilies. Include mucinases.	Viborg et al., 2019; Labourel et al., 2023
GH29 (2/2)	Eight different known activities targeting fucose containing substrates, including α-1,4-L-fucosidase	Mostly bacterial and some few fungal enzymes. Include mucinases.	Lombard et al., 2014; Labourel et al., 2023
GH30 (1/16)	14 different known activities, including β-1,6-glucosidase and Endo-β-1,4-xylanase	Mostly bacterial and some fungal enzymes. Classified in 10 subfamilies	Lombard et al., 2014

GH31 (17/17)	17 different known activities including α-galactosidase and α-mannosidase	Mostly bacterial and a few fungal enzymes. Classified in 20 subfamilies.
		Include mucinases.
			Lombard et al., 2014;
			Arumapperuma et al., 2023;
			Labourel et al., 2023

GH33 (1/2)	Five different known activities including sialidases	Mostly bacterial and a few fungal enzymes. Include mucinases.	Lombard et al., 2014; Labourel et al., 2023
GH47 (24/27)	Mannosyl-oligosaccharide α-1,2-mannosidase	Mostly eukaryotic and a few bacterial enzymes	Lombard et al., 2014
GH77 (7/14)	4-α-glucanotransferase/amylomaltase	Mostly bacterial and a few eukaryotic and archaeal enzymes	Lombard et al., 2014
GH163 (2/10)	Endo-like activity targeting the β-GlcNAc–mannose from complex N-glycans	Mostly bacterial enzymes. Could target mammalian glycoproteins, including immunoglobulins such as secretory IgA	Briliute et al., 2019

*Values in brackets (x/y) are the number of annotated genes encoding listed GH families for respectively *T. tenax* and *T. vaginalis*.

TABLE 4 Selected characteristics of the two GH families restricted to *T. vaginalis*.

GH family #*	Known enzymatic activities(s)	Comments	References
GH5_4 (1 entry)	30 different known activities. GH5 subfamily 4: endo-β-1,4-glucanases, licheninases, and xylanases	Classified in 57 subfamilies GH5 subfamily 4: typically, extracellular bacterial enzymes, also found among some ciliates and fungi.	Aspeborg et al., 2012
GH32 (1 entry)	15 different known activities on a broad range of fructose-containing substrates	Mostly bacterial and a few fungal enzymes	Lombard et al., 2014

TABLE 5 Selected annotated peptidase families representing candidate virulence factors of *T. tenax*.

Clan	Family	Number	Example from family	Candidate virulence factors*
CA	C1	46	Papain	Figueroa-Angulo et al., 2012
CA	C40	7	NlpC/P60	Barnett et al., 2023
MA	M8	36	Leishmanolysin/GP63	Ma et al., 2011; Figueroa-Angulo et al., 2012
MA	M60	13	Enhancin	Hirt et al., 2011; Nakjang et al., 2012; Riestra et al., 2022
SB	S8	31	Subtilisin	Hernández-Romano et al., 2010

*Listed references are either reviews or original publications describing the listed peptidases as virulence factors. These were either directly demonstrated to degrade host proteins or damage host cells (C1 peptidases) or through indirect mechanisms by targeting members of the microbiota (C40 peptidases). M60-like peptidases are speculated to represent virulence factors based on their cell surface location in *T. vaginalis* and enriched distribution among mammalian host associated microbial pathogens and members of the microbiota. S8 peptidases are speculated to represent virulence factors based on their cell surface location in *T. vaginalis* and known targets among several microbial pathogens including parasites, fungi and bacteria.

There is also evidence for at least 14 *T. tenax* BAP-like proteins related to TVAG_244130 (de Miguel et al., 2010), with extracellular domains that are shared with bacterial proteins (Table 6). Two of the TtBAP-like proteins have similar structural organization with eight TvBAP-like proteins (>25% identify level, similar protein length and shared TMD-CT), including the TvBAP-like proteins TVAG_166850 and TVAG_244130 (Figure 3).

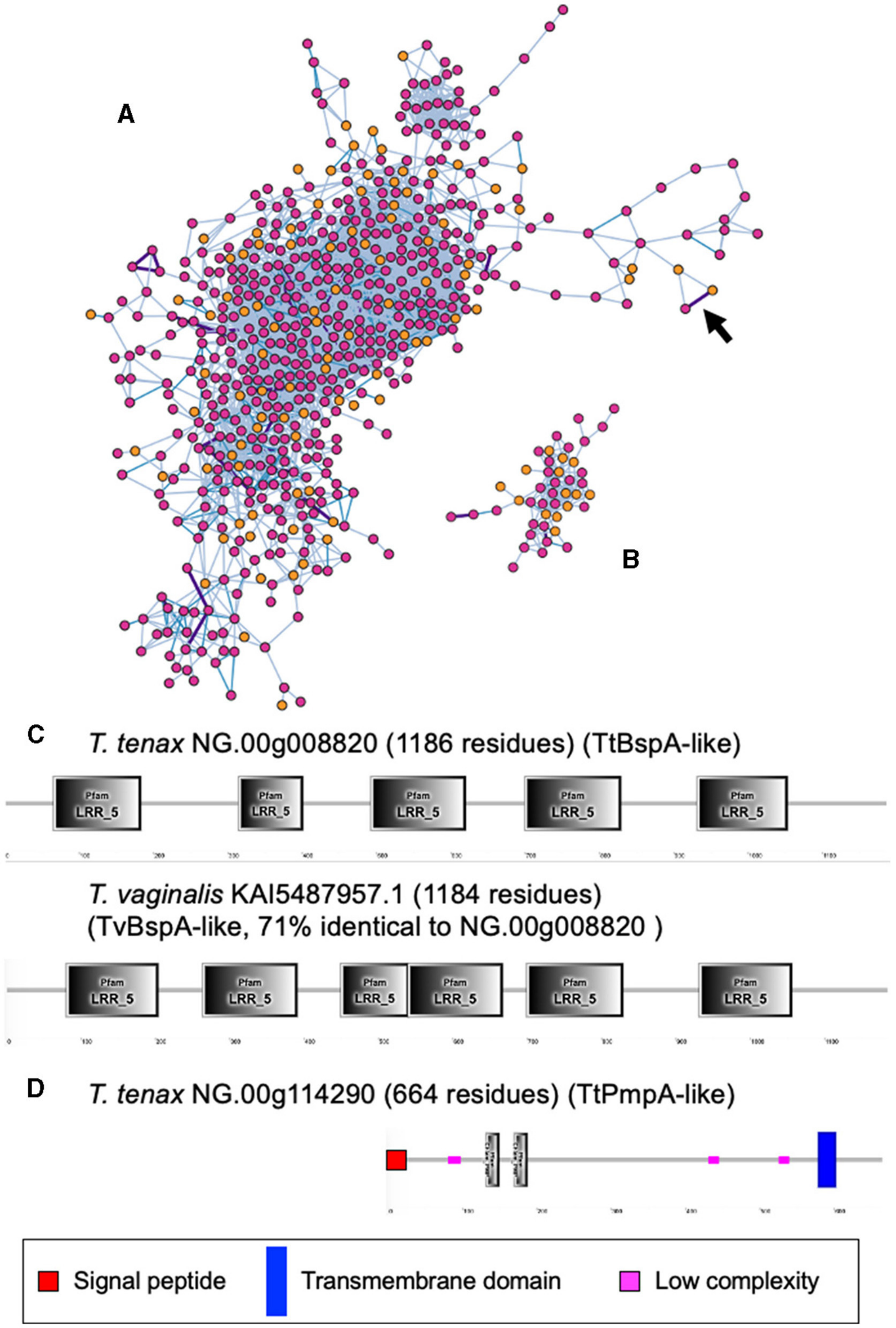

A
B
C
T. tenax NG.00g008820 (1186 residues) (TtBspA-like)
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
T. vaginalis KAI5487957.1 (1184 residues)
(TvBspA-like, 71% identical to NG.00g008820)
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
Pfam LRR_5
D
T. tenax NG.00g114290 (664 residues) (TtPmpA-like)
🟥 Signal peptide
Transmembrane domain
Low complexity

FIGURE 1
Sequence comparisons between *T. tenax* and *T. vaginalis* BspA- and Pmp-like proteins. The gene networks illustrate the relationships between **(A)** BspA- and **(B)** Pmp-like protein families. Gene networks in which nodes represent individual homologous genes and edges represent significant BLASTP alignments. Nodes are colored according to species (pink: *T. vaginalis*, orange: *T. tenax*). Edge color is scaled according to BLASTP alignment percentage identity from 25–40% (light blue), 40–70% (blue) and >70% (purple). Purple edges are displayed with greater width to aid visibility. The two illustrated networks correspond to the one with the largest number of nodes (genes) for the respective gene families. The arrow in **(A)** indicates the edge between the two BspA-like sequences with the highest level of sequence similarity between *T. tenax* and *T. vaginalis* proteins **(C)**.

(Continued)

FIGURE 1 (Continued)
(C) The structural organization inferred by SMART for the two most similar BspA-like proteins (71% identity) between *T. tenax* and *T. vaginalis* are illustrated and drawn to scale. Not all domains identified by SMART are illustrated due to overlap between some of these inferred domains. None of these proteins have an inferred SP or TMD. **(D)** The structural organization inferred by SMART for one *T. tenax* Pmp-like protein, which is inferred to possess both a SP and a TMD, with the segment with the Pmp domain (motif GGA[ILV] and FXXN) to be inferred to be exposed to the extracellular environment.

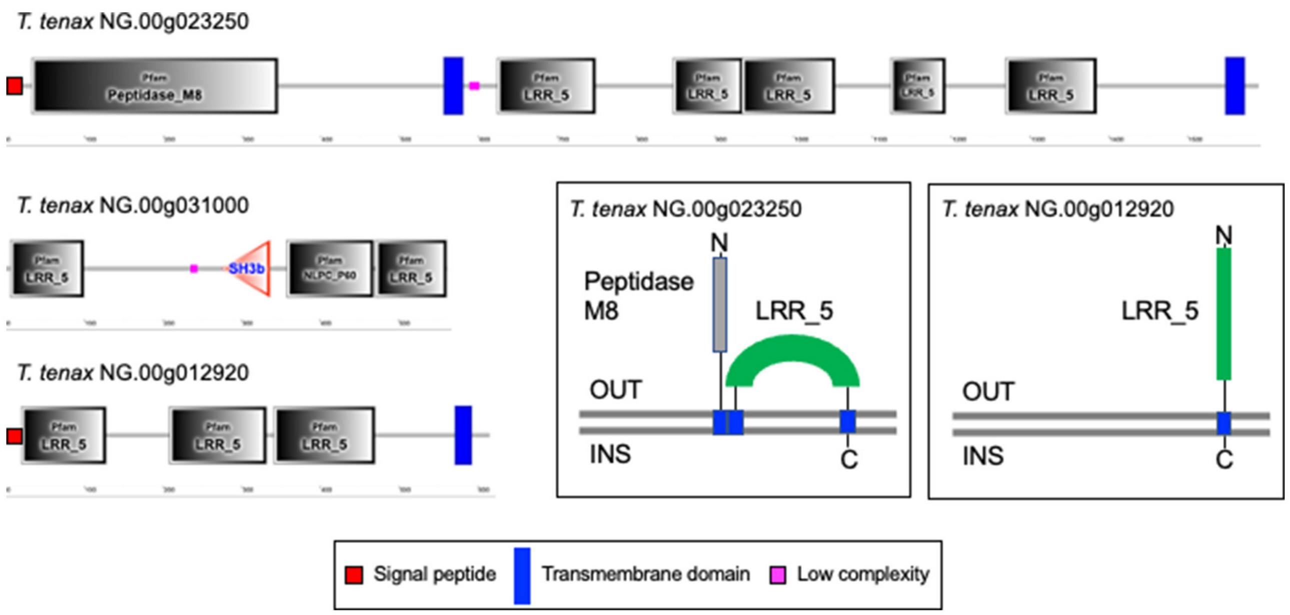


FIGURE 2
Examples of TtBspA-like proteins. The graphs represent SMART outputs drawn to scale. Two illustrated entries were selected as they possess a peptidase domain: respectively a metallopeptidase M8, leshmaniolysin domain, and a cysteine peptidase C40, NlpC/P60 domain. The entry with the metallopeptidase M8 domain has two (TMHMM—SMART output, illustrated) or three (DeepTMHMM output, see box) inferred TMDs. The topological configuration inferred from DeepTMHMM has the peptidaseM8 domain and the LRR domain to be extracellular, as illustrated within the box. These are not drawn to scale and only illustrate protein topology in relation to the bilayer of the plasma membrane of *T. tenax*. DeepTMHMM also inferred a SP, which was not identified by SMART. The C40 NlpC/P60 peptidase possess the key catalytic triad (C-H-H) and one inferred SH3b domain, as identified for other *Trichomonas* NlpC/P60 (Barnett et al., 2023). The protein *T. tenax* NG.00g012920 has the more common and characteristic structural organization and topology of BspA-like proteins among *Trichomonas* and relatives (Noël et al., 2010; Handrich et al., 2019) and possess both a SP and a TMD. The LRR domain is inferred to be exposed to the extracellular environment, consistent with a potential binding role of the LRR domain to one or more extracellular ligands, such as molecules on the host epithelial or bacteria cell surface or a specific soluble molecule. OUT, outside the cell; INS, inside the cell. The symbols for SP, TMD and low complexity sequences are indicated.

TABLE 6 *T. tenax* encodes BAP-like homologs.

Taxa	Habitat and relationship with host	Top bit score (locus tag)	Top e-value	Top hit protein length	Total number of hits
Trichomonas vaginalis G3	Urogenital tract-pathobiont (human)	348 (TVAG_335250)	$6e^{-105}$	751	60
Trichomonas tenax Hs-4:NIH	Oral cavity- pathobiont? (human, dog, cat)	303 (NG.00g068630)	$3e^{-92}$	728	14
Clostridium sp. CAG:273	Gut-normal flora (human)	102 (BN581_00590)	$7e^{-18}$	1401	1
Clostridia bacterium	Gut-normal flora (human)	99.8 (UIT70_01140)	$3e^{-17}$	652	1
Eubacterium sp.	Gut-normal flora (mouse)	100 (K2K71_04565)	$4e^{-17}$	1027	1
Clostridium sp. CAG:245	Gut-normal flora (human)	98.6 (BN559_00332)	$1e^{-16}$	893	1
Clostridium sp. CAG:245_30_32	Gut-normal flora (human)	95.1 (BHW09_08500)	$1e^{-15}$	893	1

The shown bit scores, e-values and number of hits are derived from BLASTP searchers (e-value $\leq e^{-10}$) at the NCBI Blast server using the *T. vaginalis* BAP-like protein TVAG_244130 sequence as query against the (i) nr database or (ii) the 20,286 annotated proteins from *T. tenax* Hs-4:NIH. Only the features of the top hits are listed for each taxon. The ranking, from top to bottom of the table, is according to the bit score, with the highest value on the top.

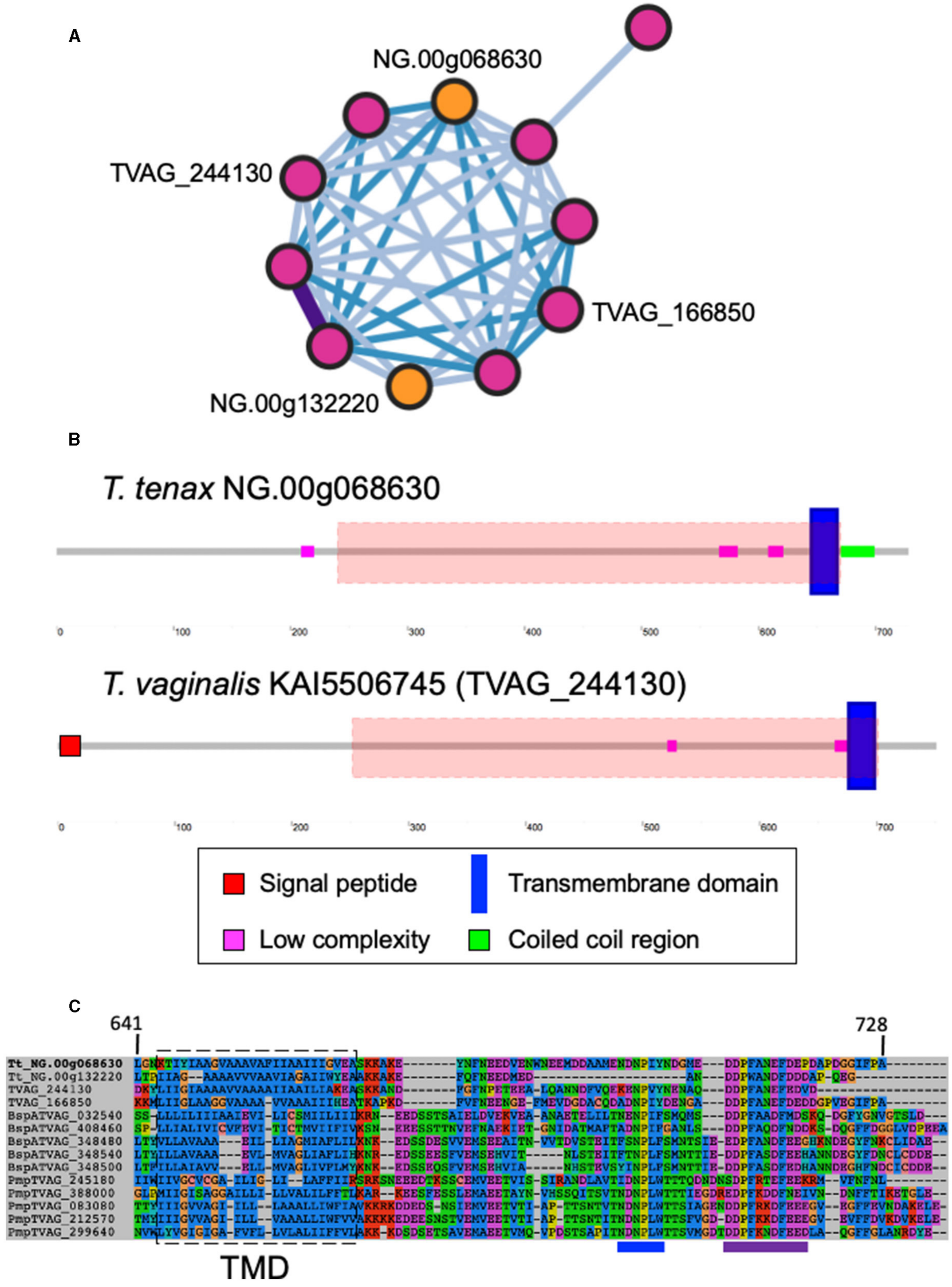


FIGURE 3

Sequence comparisons between *T. tenax* and *T. vaginalis* BAP-like proteins. **(A)** Gene network in which nodes represent individual homologous genes and edges represent significant BLASTP alignments. Nodes are colored according to species (pink: *T. vaginalis*, orange: *T. tenax*). Edge color is scaled according to BLASTP alignment percentage identity from 25-40% (light blue), 40-70% (blue) and >70% (purple). Purple edges are displayed with greater width to aid visibility. The locus tags for the two TtBAP-like annotated proteins member of the illustrated network are indicated as are the two *T. vaginalis* BAP-like proteins (2007 locus tag) with functional data (see main text). **(B)** The structural organization inferred by SMART for one *T. tenax* and one *T. vaginalis* BAP-like are illustrated and drawn to scale. The symbols for SP, TMD, low complexity and coiled coil sequences are

(Continued)

FIGURE 3 (Continued)
indicated. The red dashed boxes illustrate the position of the HHPred top hit alignments between the alignment of 13 selected BAP-like proteins used as query: "Cell surface glycoprotein; S-layer csg, STRUCTURAL PROTEIN; HET: CA, BGC; 3.46A {Haloferax volcanii DS2}", probability: 98.34, e-values: 0.0053. The alignment starts and ends at position 222 and 668, respectively, from NG.00g06863. **(C)** Multi sequence alignment illustrating the conservation of the TMD and cytoplasmic tail for selected BAP-like, BspA and Pmp-like proteins. The approximate position of the inferred TMD is indicated by the dashed box and the position of the [FY]NPX[YF]-like motif and the acidic cluster are indicated by respectively a blue and purple line below the alignment, these could represent signal for endocytosis (see Hirt et al., 2011; Handrich et al., 2019).

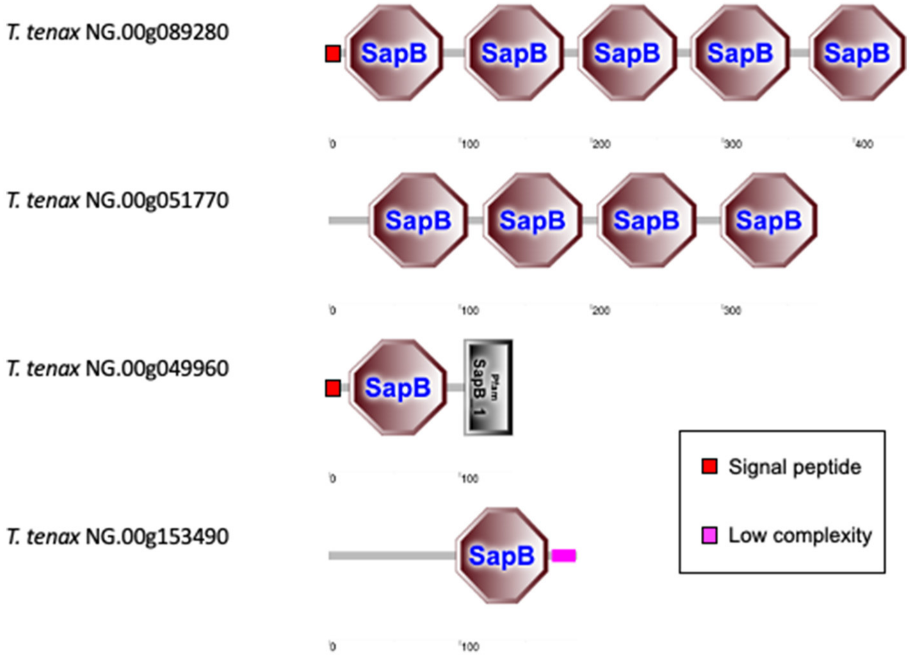


FIGURE 4
Examples of TtSAPLIP-like protein structural diversity. The graphs represent SMART outputs drawn to scale. The four illustrated examples focus on the proteins with multiple SAPLIP domains (five and four) and two examples of proteins with two or one inferred SAPLIP domains only. The remaining six TtSAPLIP-like proteins have only one inferred SAPLIP. The symbols for SP and low complexity sequences are indicated.

3.3.3 Saposin-like proteins

The *T. vaginalis* G3_2007 annotation identified 12 candidate pore-forming effectors that belong to the conserved family of saposin-like proteins (SAPLIP) (named TvSAPLIP1-12) that represent important candidate virulence factors (Carlton et al., 2007; Hirt et al., 2011; Diaz et al., 2020). We identified 10 SAPLIP *T. tenax* encoding genes (Supplementary Table 7). Three TtSAPLIP proteins display multiple domains with five, four and two different SAPLIP domains (Figure 4). The presence of a SP was inferred for seven TtSAPLIP, consistent with extracellular functions (Supplementary Table 7, Figure 4).

3.3.4 Extracellular vesicles

As carefully reviewed by Riestra et al. (2022), there are published evidence, including electron microscopy, proteomics surveys and cell biology, demonstrating that *T. vaginalis* can secrete two types of extracellular vesicles (EVs), exosomes and microvesicles (MVs) and that these modulate several important aspects of *T. vaginalis*-host cells interactions (Twu et al., 2013; Rai and Johnson, 2019; Artuyants et al., 2020; Molgora et al., 2023; Salas et al., 2023). We identified *T. tenax* homologs for the Endosomal Sorting Complex Required for Transport (ESCRT) machinery made of four complexes (complexes I to IV) (Leung et al., 2008), which are involved in exosome formation in multi-vesicle bodies (MVBs) and secretion in many eukaryotes, strongly suggesting that *T. tenax* also possess MVBs and could secrete exosomes (Supplementary Table 8).

4 Discussion

4.1 CAZymes

CAZymes, including GHs, are of great interest as their glycans substrates are widely distributed across all cellular life forms as well as viruses. Glycans mediate a myriad of biological functions, including handling of carbon reserves, structural roles, or as mediators of intra- and intercellular interactions between organisms or organisms and viruses. CAZymes and GH in particular, but also CBM-containing proteins, include virulence factors of numerous cellular and viral pathogens (Garron and Henrissat, 2019; Wardman et al., 2022). A diverse array of

GH families was identified in the two *Trichomonas* species and the majority of these are shared between them, suggesting that these play key roles in the biology of both the oral cavity and the urogenital tract for *T. tenax* and *T. vaginalis*, respectively. A number of families are characterized by relatively larger membership with more than 10 members in one or both species. Three families, GH13, GH31 and GH47 stand out as making up the largest families present in both species (ranging from 14 to 27 entries), further emphasizing their potential importance in the mucosal lifestyle of the respective species.

Notably, some GH entries represent candidate virulence factors. Seven GH families (GH2, GH16, GH18, GH16, GH29, GH33 and GH163) could include members from either species that target mucins (Labourel et al., 2023) (candidate endo-β-N-acetylglucosaminidases in GH163, candidate fucosidases in GH29, candidate sialidases in GH33, candidate β-galactosidases in GH2), important structural proteins at mucosal surfaces mediating key innate immune functions in host-microbe interactions (Hansson, 2020). *T. vaginalis* might be able to degrade mucins more extensively as it has candidate O-glycan endo-β-1,4-galactosidases in GH16, while *T. tenax* does not seem to encode a homolog with the same predicted activity. Mucins are typically targeted by mucosal pathogens as well as some members of the microbiota (Labourel et al., 2023). One or more of these enzymes could also target SIgA, a key immunoglobulin in the oral cavity (Feller et al., 2013) and the urogenital tract (Garcia et al., 2015), an activity demonstrated for a GH163 from a common member of the human gut microbiota (Briliute et al., 2019).

Several GH families could also target members of the microbiota cell walls, either of bacteria (lysozymes GH19 and GH25) or fungi (chitinase GH19) or GH47, which target glycoproteins containing α-1,2-linked mannose residues. Candidate lysozymes could be functional partners of the NlpC/P60 peptidases, which were demonstrated to target peptidoglycans (Barnett et al., 2023), to contribute to deconstruct bacterial cell walls. These sets of enzymes could be important for the two *Trichomonas* species to target bacteria for interspecies competition and extraction of important source of nutrients. As fragments of peptidoglycans represent strong pro-inflammatory molecules (Irazoki et al., 2019), the products of the combined activities of these enzymes (GH and peptidases) could also be an important factor contributing to inflammation, such as the damaging inflammations characteristic of both periodontitis (Baker et al., 2024) and trichomoniasis (Mercer and Johnson, 2018).

As most of these GH families are known to mediate multiple distinct enzymatic activities, it will be important to study their exact substrate specificities, expression profile and cell biology to experimentally establish their functional importance for *Trichomonas* species.

4.2 Candidate surface proteins and exosomes

Several key aspects of host-microbe interactions are mediated by cell surface proteins and EVs, such as exosomes, which modulate cell-cell interactions and contact-dependent cytolysis (e.g., Hirt et al., 2011; Riestra et al., 2022). As for *T. vaginalis* (Carlton et al., 2007; Hirt et al., 2011; Handrich et al., 2019; Riestra et al., 2022), the annotation of *T. tenax* proteins identified a large number of candidate surface proteins including members of TtBspA-like, TtPmp-like and TtBAP-like families. Two TtBspA-like proteins possess a peptidase domain (one with an M8 metallopeptidase domain and the other with NlpC/P60 cysteine peptidase domain), a configuration distinct from the single TvBspA-like protein with a peptidase domain (a partial serine peptidase S8/S53-like domain) (Noël et al., 2010). Potential surface proteins also include several candidate peptidases and GHs, as they possess inferred TMDs. There is transcriptional evidence for the expression of 81 TtBspA-like and 11 TtPmp-like encoding genes (Handrich et al., 2019). In *T. vaginalis*, one or more members of these three families have been experimentally demonstrated to be expressed on the cell surface and to mediate interactions with vaginal epithelial cells (de Miguel et al., 2010; Handrich et al., 2019). Hence some of the identified homologs from these protein families in *T. tenax* could contribute to interactions of this species with epithelial cells of the oral cavity, interactions which are likely key to mediating the long-term colonization of the periodontal pocket, and that could also contribute to cell contact-mediated cytolysis of host cells (Ribeiro et al., 2015; Matthew et al., 2023). Our gene network analyses that used stringent coverage of the pairwise alignment length criteria (≥90% coverage) indicate that a number of BspA-, Pmp and BAP-like entries have similar structural organization, with a number of these sharing similar length and a TMD and cytoplasmic tail (TMD-CT). Visual inspection of the cytoplasmic tail of TtBAP- and TvBAP-like proteins identified conserved TMD-CT shared with some TvBspA- and TvPmp-like proteins (Hirt et al., 2011; Handrich et al., 2019). This indicates that the gene fusion events underlying these shared TMD-CT domains between BspA and Pmp-like coding genes (Hirt et al., 2011) have also involved some members of the BAP-like family (Handrich et al., 2019) and that these fusion events might have occured in a common ancestor to *T. tenax* and *T. vaginalis*. This further supports the functional importance of this type of TMD-CT and in the context of various unrelated extracellular domains: (i) BspA-like proteins characterized by a specific type of leucine-rich repeat (TpLRR), (ii) Pmp-like repeats (that include two specific motifs GGA[ILV] and FXXN) and (iii) BAP domain (related to "Cell surface glycoprotein; S-layer"). The two TvBAP-like proteins (TVAG_166850, TVAG_244130) were shown to be significantly more highly expressed by more adherent *T. vaginalis* strains compared to less adherent strains (de Miguel et al., 2010). Overexpression of some members of these three gene families, one TvBspA, two TvPmp (Handrich et al., 2019) and two BAP-like proteins (TVAG_166850, TVAG_244130) (de Miguel et al., 2010) significantly increased *T. vaginalis* binding to vaginal epithelial cells in *in vitro* binding assays, suggesting that some members of the TtBspA-like, TtPmp-like and TtBAP-like proteins families could contribute to *T. tenax* binding to oral epithelial cells. The Tv/TtBspA-like proteins could also mediate interactions between the *Trichomonas* species and members of the microbiota based on their known functions in bacteria-bacteria interactions among oral bacterial pathogens (Sharma, 2010), which could contribute to bacteria phagocytosis, as previously speculated for *T. vaginalis* (Noël et al., 2010). Notably, the TtBspA-like proteins could also contribute to the inflammatory response of the periodontal pocket, as demonstrated for some oral bacterial BspA proteins (Sharma, 2010).

Some members of these protein families also contribute to the proteome of *T. vaginalis* exosomes (Riestra et al., 2022; Ong et al., 2024). Our identification of homologs of subunits of the ESCRT machinery (with homologs identified for all four complexes I to IV), suggests that exosomes are likely produced and secreted by *T. tenax*. Hence it will be important to affirm the presence and functional characteristic of EVs in *T. tenax*, including their protein and RNA complement as has been done for *T. vaginalis* (Riestra et al., 2022; Matthew et al., 2023; Ong et al., 2024). These potentially include members of the BspA-like proteins and peptidases (e.g. some calpain-like cysteine peptidase and metallopeptidases, see next section). These different proteins possibly play multiple roles, including modulating *T. tenax* binding properties to host cells, and mediating cytolytic capabilities, which could both contribute to the pathobiology of periodontitis. Other proteins identified in *T. vaginalis* exosomes include membrane protein tetraspanins and CBM20-GH77 enzymes (candidate 4-α-glucanotransferase) (Rai and Johnson, 2019). We identified homologs of these proteins in *T. tenax*, including 15 tetraspanins of which 10 were inferred to possess the canonical four TMDs of tetraspanins (Supplementary Table 2) and five proteins with the CBM20-GH77 domain configuration, with one protein inferred to possess a TMD (Supplementary Table 3). In *T. vaginalis* some tetraspanin proteins were also shown to mediate *T. vaginalis* binding to self (Cóceres et al., 2015), thus contributing to swarming, which is thought to have implication for its virulence (Hirt, 2013). Notably, the CBM20 domain from one of the *T. vaginalis* GH77 exosomal proteins was demonstrated to bind to heparan sulfate from host proteoglycans (Rai and Johnson, 2019). Hence these different proteins represent primary targets to study *T. tenax* interactions with oral epithelial cells and with itself. Cell surface proteins and EVs could also play important roles in *T. vaginalis* and *T. tenax* interactions with immunocytes and members of the microbiota or extracellular matrix (ECM) proteins.

4.3 Candidate peptidases and pore-forming proteins

Peptidases are considered to mediate and modulate important aspects of *T. vaginalis* cytoadherence and host-cell cytolysis, and to degrade specific host proteins including immunoglobulins, the complement, ECM proteins and potentially some host-produced anti-microbial peptides (Carlton et al., 2007; Hirt et al., 2011; Figueroa-Angulo et al., 2012; Fichorova et al., 2013; Riestra et al., 2022).

Hence these enzymes are of primary interest to investigate the molecular basis of *T. tenax* pathobiology. Indeed, initial investigations indicated the secretion of cysteine peptidases and metallopeptidases by *T. tenax* (reviewed in Matthew et al., 2023). Like *T. vaginalis* (Carlton et al., 2007; Hirt et al., 2011), *T. tenax* is endowed with a large number and broad diversity of peptidases. A number of these could be either secreted as soluble proteins, secreted as exosomes cargo, or expressed on the cell surface or within organelles such as the lysosomes, where they could contribute to the degradation of proteins on the cell surface of the parasite or from phagocytosed microbial or human cells. Several M8 peptidases have an inferred SP and/or one or more TMD. One member of the M8 family was characterized with an LRR BspA-like containing domain and inferred to possess three TMDs and exposed to either the cell surface or lumen of an organelle, a location consistent with the function of the related M8 peptidases, including Leishmanolysin, an important virulence factor among *Leishmania* species (Hirt et al., 2011).

In addition to peptidases potentially targeting host and microbial proteins, some peptidases are also likely to specifically target peptides of bacterial peptidoglycans. These include the NlpC/P60 peptidases, and initial analyses of the first published draft genome of *T. tenax* (Benabdelkader et al., 2019) identified six genes encoding bacteria-like TtNlpC/P60 peptidases (Barnett et al., 2023). The genes encoding these NlpC/P60 peptidases were likely acquired by two distinct lateral gene transfers events from distinct gram-positive bacteria donors into a common ancestor of *Trichomonas* species as these are found in the genomes of *T. vaginalis*, *T. gallinae* (Alrefaei et al., 2019) and *T. tenax* (Barnett et al., 2023). A total of seven distinct TtNlpC/P60 encoding genes were identified when annotating the more recent draft genome (Yang et al., 2022). The latest identified TtNlpC/P60 protein is characterized by a distinct structural configuration, which include two segments of LRR BspA-like proteins, one on each end of the protein. As the LRR of some cell surface BspA-like proteins from oral bacteria mediate bacteria-bacteria interactions (Sharma, 2010), these two LRR segments could target this peptidase to the cell surface of specific bacteria. Six of the seven TtNlpC/P60 have inferred SP, consistent with digesting extracellular targets, such as the peptidoglycans from bacteria. The identification of several candidate GHs potentially targeting the glycans from peptidoglycans (e.g., TtGH25, candidate lysozymes) suggests that the TtNlpC/P60 protein could be functional partners of these CAZymes to effectively deconstruct peptidoglycans from the bacterial members of the oral and vaginal bacterial microbiota. As peptidoglycan fragments derived from the combined activity of peptidases and CAZymes are very strong pro-inflammatory molecules (Irazoki et al., 2019), these enzymes could represent important indirect virulent factors contributing to inducing damaging inflammations characteristic of periodontitis (Baker et al., 2024) and trichomoniasis (Mercer and Johnson, 2018; Riestra et al., 2022). In addition, as *T. vaginalis* and *T. tenax* are associated with dysbiotic microbiota, the potential preferential targeting of mutualistic bacteria by these enzymes, such as *Lactobacillus* species (that have anti-inflammatory properties and other important protective functions) typically affected/eradicated by *T. vaginalis* infections, these enzymes could also contribute to the worsening of the inflammatory response of the oral and vaginal mucosa.

The 10 TtSAPLIP candidates pore-forming proteins are also potentially important virulent factors for *T. tenax* as discussed for *T. vaginalis* (Hirt et al., 2011; Diaz et al., 2020). The functional characterization of the TvSAPLIP12 demonstrated that it is haemolytic and cytotoxic toward epithelial cells and the bacterium *Escherichia coli* (Diaz et al., 2020). The inference of SP for the majority of the TtSAPLIP proteins is consistent with the hypothesis that these candidate virulence factors are secreted and able to act as pore–forming proteins targeting host cells and/or members of the microbiota as shown (or suggested) for the TvSAPLIP proteins (Hirt et al., 2011; Diaz et al., 2020; Margarita et al., 2022).

5 Conclusions

Our detailed integrated annotations of the two *T. tenax* draft genomes from the currently single strain available from the American Type Culture Collections (ATCC, strain Hs-4:NIH) further support a model where *T. tenax* is a pathobiont that directly and/or indirectly contribute to periodontitis (Ribeiro et al., 2015; Marty et al., 2017; Benabdelkader et al., 2019; Bisson et al., 2019; Matthew et al., 2023). Comparison between *T. tenax* and *T. vaginalis* annotations also identified some novel shared candidate virulence factors including GHs potentially targeting host or bacterial glycans, making them relevant in both species, and warranting further study in both species. Based on the features of the protein complement of *T. tenax* shared with *T. vaginalis*, the oral protist could contribute to worsening the prognosis and/or initiating periodontitis in multiple ways including through host cell cytolysis and bacteria targeting, which together could contribute synergistically at inducing and/or amplifying damaging inflammations. Periodontitis is primarily recognized as a condition driven by bacterial dysbiosis, which is sufficient to develop the condition (Baker et al., 2024). However, the high prevalence of *T. tenax* among humans and pets (dogs and cats) and the heterogeneity of disease prognosis between individuals makes it potentially highly relevant to numerous cases of periodontitis and could help also at stratifying patients (Marty et al., 2017). Hence, affected hosts could potentially benefit from treatments targeting both *T. tenax* and pathogenic bacteria to reduce/eliminate the damaging inflammation characteristic of periodontitis. Establishing the potential causality of *T. tenax* for periodontitis will require further investigation, including longitudinal studies among humans and animals, and the molecular and cellular characterization of several candidate virulence factors. A notable virulence factor associated with *T. vaginalis* are the *Trichomonas vaginalis* viruses (TVV), which are members of the totoviridae known to have strong pro-inflammatory properties (Fichorova et al., 2013). Hence, it will be also very interesting to investigate the potential presence and relevance of related TVV-like viruses in *T. tenax* isolated from humans and pets. Furthermore, *T. vaginalis* is associated with two mycoplasma-like endosymbiotic bacteria, which are known to contribute to promoting *T. vaginalis* growth and to boosting virulence by up-regulating adhesion to human epithelial cells and haemolysis in *in vitro* conditions (Margarita et al., 2022). It is currently unknown if sets of related endosymbionts are associated with *T. tenax*. It will be in particular important to sequence genomes from across the spectrum of the known genetic diversity of *T. tenax*, which includes specific genotypes associated with worse periodontitis (Benabdelkader et al., 2019). This should include recent clinical isolates from both humans and animals (e.g. dogs and cats), as the existing strain (Hs-4:NIH) has been grown *in vitro* since 1959, and thus may not represent the best, albeit a very valuable and important, model system to study *T. tenax*-host interactions.

Data availability statement

The original draft genome sequence data analyzed in this study are available at the NCBI (NCBI Genome assemblies PRJEB22701 and ASM2309173v1). All the key outputs of the sequence analyses for this study can be found in the Supplementary material. They are all described in the main text and in the Supplementary material.

Author contributions

LM: Conceptualization, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. AH: Data curation, Formal analysis, Investigation, Writing – review & editing. NB: Formal analysis, Investigation, Visualization, Writing – review & editing. JC: Data curation, Funding acquisition, Methodology, Project administration, Writing – review & editing, Resources. BH: Data curation, Formal analysis, Methodology, Writing – review & editing, Resources. SS: Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – review & editing, Resources. RH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing, Resources.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Commonwealth Scholarship Commission in the UK (grant number CMCS-2019-109, LM PhD student and RH supervisor), Newcastle University and the Biotechnology and Bioscience Research Council Doctoral Training Partnership for Newcastle, Liverpool and Durham (grant numbers BB/T008695/1, AH PhD student, RH supervisor and BB/M011186/1, NB PhD student, RH supervisor). Research reported in this publication was also supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R21AI149449 (JC).

Acknowledgments

This research made use of the Rocket High Performance Computing service at Newcastle University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1437572/full#supplementary-material

SUPPLEMENTARY DATA 1
The GFF3 file for the annotation of *T. tenax* sequence data from Benabdelkader et al. (2019) (zipped).

SUPPLEMENTARY DATA 2
The GFF3 file for the annotation of *T. tenax* sequence data from Yang et al. (2022) (zipped).

SUPPLEMENTARY DATA 3
The FASTA file with all 20,286 annotated proteins (integrated from the above two annotations) (flat text file).

SUPPLEMENTARY TABLE 1
T. vaginalis G3 look up table integrating and linking the 2007 and 2022 annotations (excel file).

SUPPLEMENTARY TABLE 2
T. tenax 20,286 locus tags for the protein coding genes, their annotation, protein length, TMHMM, SignalP and Phobius inferences and the protein sequence (excel table).

SUPPLEMENTARY TABLE 3
The CAZy annotation for both *T. tenax* and *T. vaginalis* G3_2007 (excel table).

SUPPLEMENTARY TABLE 4
The annotated peptidases for *T. tenax* (excel table).

SUPPLEMENTARY TABLE 5
The annotated BspA-like proteins for *T. tenax* (excel table).

SUPPLEMENTARY TABLE 6
The annotated Pmp-like proteins for *T. tenax* (excel table).

SUPPLEMENTARY TABLE 7
The annotated SAPLIP proteins for *T. tenax* (excel table).

SUPPLEMENTARY TABLE 8
The annotated ESCRT subunits for complex I-IV for *T. tenax* (excel table).

References

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10. doi: 10.1093/nar/gkw343

Alrefaei, A. F., Low, R., Hall, N., Jardim, R., Dávila, A., Gerhold, R., et al. (2019). Multilocus analysis resolves the European finch epidemic strain of *Trichomonas gallinae* and suggests introgression from divergent trichomonads. *Genome Biol. Evol.* 11, 2391–2402. doi: 10.1093/gbe/evz164

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Artuyants, A., Campos, T. L., Rai, A. K., Johnson, P. J., Dauros-Singorenko, P., Phillips, A., et al. (2020). Extracellular vesicles produced by the protozoan parasite *Trichomonas vaginalis* contain a preferential cargo of tRNA-derived small RNAs. *Int. J. Parasitol.* 50, 1145–1155. doi: 10.1016/j.ijpara.2020.07.003

Arumapperuma, T., Li, J., Hornung, B., Soler, N. M., Goddard-Borger, E. D., Terrapon, N., et al. (2023). A subfamily classification to choreograph the diverse activities within glycoside hydrolase family 31. *J. Biol. Chem.* 299:103038. doi: 10.1016/j.jbc.2023.103038

Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H., and Henrissat, B. (2012). Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* 12, 1–16. doi: 10.1186/1471-2148-12-186

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Carlton, J. M., Dommer, J., Fischer, S., et al. (2009). GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.* 37, D526–D530. doi: 10.1093/nar/gkn631

Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S., and He, X. (2024). The oral microbiome: diversity, biogeography and human health. *Nat. Rev. Microbiol.* 22, 89–104. doi: 10.1038/s41579-023-00963-6

Barnett, M. J., Pinheiro, J., Keown, J. R., Biboy, J., Gray, J., Lucinescu, I. W., et al. (2023). NlpC/P60 peptidoglycan hydrolases of *Trichomonas vaginalis* have complementary activities that empower the protozoan to control host-protective lactobacilli. *PLoS Pathog.* 19:e1011563. doi: 10.1371/journal.ppat.1011563

Barrett, A. J., and Rawlings, N. D. (2001). Evolutionary lines of cysteine peptidases. *Biol. Chem.* 382, 727–734. doi: 10.1515/bchm.2001.382.5.727

Benabdelkader, S., Andreani, J., Gillet, A., Terrer, E., Pignoly, M., Chaudet, H., et al. (2019). Specific clones of *Trichomonas tenax* are associated with periodontitis. *PLoS ONE* 14:e0213338. doi: 10.1371/journal.pone.0213338

Bisson, C., Dridi, S. M., and Machouart, M. (2019). Assessment of the role of *Trichomonas tenax* in the etiopathogenesis of human periodontitis: a systematic review. *PLoS ONE* 14:e0226266. doi: 10.1371/journal.pone.0226266

Briliute, J., Urbanowicz, P. A., Luis, A. S., Basl,é, A., Paterson, N., Rebello, O., et al. (2019). Complex N-glycan breakdown by gut *Bacteroides* involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nature Microbiol.* 4, 1571–1581. doi: 10.1038/s41564-019-0466-x

Brosh-Nissimov, T., Hindiyeh, M., Azar, R., Smollan, G., Belausov, N., Mandelboim, M., et al. (2019). A false-positive *Trichomonas vaginalis* result due to *Trichomonas tenax* presence in clinical specimens may reveal a possible *T. tenax* urogenital infection. *Clini. Microbiol. Infect.* 25, 123–124. doi: 10.1016/j.cmi.2018.09.011

Buchfink, B., Reuter, K., and Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformat.* 10, 1–9. doi: 10.1186/1471-2105-10-421

Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., et al. (2007). Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315, 207–212. doi: 10.1126/science.1132894

Carlton, J. M., Malik, S. B., Sullivan, S. A., Sicheritz-Pontén, T., Tang, P., and Hirt, R. P. (2010). "The genome of *Trichomonas vaginalis*," in *Anaerobic Parasitic Protozoa: Genomics and Molecular Biology*. Wymondham: Caister Academic Press, 45–80.

Cepicka, I., Hampl, V., and Kulda, J. (2010). Critical taxonomic revision of parabasalids with description of one new genus and three new species. *Protist* 161, 400–433. doi: 10.1016/j.protis.2009.11.005

Cóceres, V. M., Alonso, A. M., Nievas, Y. R., Midlej, V., Frontera, L., Benchimol, M., et al. (2015). The C-terminal tail of tetraspanin proteins regulates their intracellular distribution in the parasite *Trichomonas vaginalis*. *Cell. Microbiol.* 17, 1217–1229. doi: 10.1111/cmi.12431 de Miguel, N., Lustig, G., Twu, O., Chattopadhyay, A., Wohlschlegel, J. A., and Johnson, P. J. (2010). Proteome analysis of the surface of Trichomonas vaginalis reveals novel proteins and strain-dependent differential expression. *Mol. Cellular Proteom.* 9, 1554–1566. doi: 10.1074/mcp.M000022-MCP201

Diaz, N., Lico, C., Capodicasa, C., Baschieri, S., Dess,ì, D., Benvenuto, E., et al. (2020). Production and functional characterization of a recombinant predicted pore-forming protein (TVSAPLIP12) of *Trichomonas vaginalis* in *Nicotiana benthamiana* Plants. *Front. Cell. Infect. Microbiol.* 10:581066. doi: 10.3389/fcimb.2020.581066

Drula, E., Garron, M. L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* 50, D571–D577. doi: 10.1093/nar/gkab1045

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Eslahi, A. V., Olfatifar, M., Abdoli, A., Houshmand, E., Johkool, M. G., Zarabadipour, M., et al. (2021). The neglected role of *Trichomonas tenax* in oral diseases: a systematic review and meta-analysis. *Acta Parasitologica* 77, 715–732. doi: 10.1007/s11686-021-00340-4

Feller, L., Altini, M., Khammissa, R. A. G., Chandran, R., Bouckaert, M., and Lemmer, J. (2013). Oral mucosal immunity. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* 116, 576–583. doi: 10.1016/j.oooo.2013.07.013

Fichorova, R. N., Buck, O. R., Yamamoto, H. S., Fashemi, T., Dawood, H. Y., Fashemi, B., et al. (2013). The villain team-up or how *Trichomonas vaginalis* and bacterial vaginosis alter innate immunity in concert. *Sex. Transm. Infect.* 89, 460–466. doi: 10.1136/sextrans-2013-051052

Figueroa-Angulo, E. E., Rendón-Gandarilla, F. J., Puente-Rivera, J., Calla-Choque, J. S., Cárdenas-Guerra, R. E., Ortega-López, J., et al. (2012). The effects of environmental factors on the virulence of *Trichomonas vaginalis*. *Microb. Infect.* 14, 1411–1427. doi: 10.1016/j.micinf.2012.09.004

Garcia, M. R., Patel, M. V., Shen, Z., Fahey, J. V., Biswas, N., Mestecky, J., et al. (2015). Mucosal immunity in the human female reproductive tract. *Mucosal Immunol.* 2015, 2097–2124. doi: 10.1016/B978-0-12-415847-4.00108-7

Garron, M. L., and Henrissat, B. (2019). The continuing expansion of CAZymes and their families. *Curr. Opin. Chem. Biol.* 53, 82–87. doi: 10.1016/j.cbpa.2019.08.004

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:r7. doi: 10.1186/gb-2008-9-1-r7

Halary, S., McInerney, J. O., Lopez, P., and Bapteste, E. (2013). EGN: a wizard for construction of gene and genome similarity networks. *BMC Evol. Biol.* 13, 1–9. doi: 10.1186/1471-2148-13-146

Hallgren, J., Tsirigos, K. D., Pedersen, M. D., Almagro Armenteros, J. J., Marcatili, P., Nielsen, H., et al. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *BioRxiv* [preprint]BioRxiv 2022–04. doi: 10.1101/2022.04.08.487609

Handrich, M. R., Garg, S. G., Sommerville, E. W., Hirt, R. P., and Gould, S. B. (2019). Characterization of the BspA and Pmp protein family of trichomonads. *Parasites Vectors* 12:406. doi: 10.1186/s13071-019-3660-z

Hansson, G. C. (2020). Mucins and the microbiome. *Annu. Rev. Biochem.* 89, 769–793. doi: 10.1146/annurev-biochem-011520-105053

Hernández-Romano, P., Hernández, R., Arroyo, R., Alderete, J. F., and Lopez-Villasenor, I. (2010). Identification and characterization of a surface-associated, subtilisin-like serine protease in *Trichomonas vaginalis*. *Parasitology* 137, 1621–1635. doi: 10.1017/S003118201000051X

Hirt, R. P. (2013). *Trichomonas vaginalis* virulence factors: an integrative overview. *Sex. Transm. Infect.* 89, 439–443. doi: 10.1136/sextrans-2013-051105

Hirt, R. P., de Miguel, N., Nakjang, S., Dessi, D., Liu, Y. C., Diaz, N., et al. (2011). *Trichomonas vaginalis* pathobiology: new insights from the genome sequence. *Adv. Parasitol.* 77, 87–140. doi: 10.1016/B978-0-12-391429-3.00006-X

Honigberg, B. M. (1990). "Trichomonads found outside the urogenital tract of humans," in *Trichomonads Parasitic in Humans* (New York, NY: Springer), 342–393.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi: 10.1093/nar/gkv1248

Humann, J. L., Lee, T., Ficklin, S. P., Cheng, C. H., Hough, H., Jung, S., et al. (2019). "GenSAS v6. 0: a web-based platform for structural and functional annotation of model and non-model organisms," in *Plant and Animal Genome XXVII Conference (January 12–16, 2019)* (PAG).

Irazoki, O., Hernandez, S. B., and Cava, F. (2019). Peptidoglycan muropeptides: release, perception, and functions as signaling molecules. *Front. Microbiol.* 10:429805. doi: 10.3389/fmicb.2019.00500

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Käll, L., Krogh, A., and Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res.* 35(Suppl_2), W429-W432. doi: 10.1093/nar/gkm256

Kanehisa, M. (2002). "The KEGG database," in *'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium, Vol. 247* (Chichester: John Wiley & Sons, Ltd.), 91–103.

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kellerová, P., and Tachezy, J. (2017). Zoonotic Trichomonas tenax and a new trichomonad species, *Trichomonas brixi* n. sp., from the oral cavities of dogs and cats. *Int. J. Parasitol.* 47, 247–255. doi: 10.1016/j.ijpara.2016.12.006

Kelly, S., and Maini, P. K. (2013). DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. *PLoS ONE* 8:e58537. doi: 10.1371/journal.pone.0058537

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:r36. doi: 10.1186/gb-2013-14-4-r36

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:59. doi: 10.1186/1471-2105-5-59

Kriventseva, E. V., Rahman, N., Espinosa, O., and Zdobnov, E. M. (2007). OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* 36(Suppl_1), D271-D275. doi: 10.1093/nar/gkm845

Labourel, A., Parrou, J. L., Deraison, C., Mercier-Bonin, M., Lajus, S., and Potocki-Veronese, G. (2023). O-Mucin-degrading carbohydrate-active enzymes and their possible implication in inflammatory bowel diseases. *Essays Biochem.* 67:331. doi: 10.1042/EBC20220153

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493–D496. doi: 10.1093/nar/gkx922

Leung, K. F., Dacks, J. B., and Field, M. C. (2008). Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage. *Traffic* 9, 1698–1716. doi: 10.1111/j.1600-0854.2008.00797.x

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 33, 6494–6506. doi: 10.1093/nar/gki937

Ma, L., Meng, Q., Cheng, W., Sung, Y., Tang, P., Hu, S., et al. (2011). Involvement of the GP63 protease in infection of Trichomonas vaginalis. *Parasitol. Res.* 109, 71–79. doi: 10.1007/s00436-010-2222-2

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Margarita, V., Bailey, N. P., Rappelli, P., Diaz, N., Dess,ì, D., Fettweis, J. M., et al. (2022). Two different species of Mycoplasma endosymbionts can influence *Trichomonas vaginalis* pathophysiology. *MBio* 13, e00918–e00922. doi: 10.1128/mbio.00918-22

Margarita, V., Congiargiu, A., Diaz, N., Fiori, P. L., and Rappelli, P. (2023). *Mycoplasma hominis* and *Candidatus Mycoplasma girerdii* in *Trichomonas vaginalis*: peaceful cohabitants or contentious roommates? *Pathogens* 12:1083. doi: 10.3390/pathogens12091083

Maritz, J. M., Land, K. M., Carlton, J. M., and Hirt, R. P. (2014). What is the importance of zoonotic trichomonads for human health? *Trends Parasitol.* 30, 333–341. doi: 10.1016/j.pt.2014.05.005

Martin-Garcia, D. F., Sallam, M., Garcia, G., and Santi-Rocca, J. (2022). Parasites in periodontal health and disease: a systematic review and meta-analysis. *Periodontitis: Adv. Exp. Res.* 2022, 95–111. doi: 10.1007/978-3-030-96881-6_5

Marty, M., Lemaitre, M., Kemoun, P., Morrier, J. J., and Monsarrat, P. (2017). *Trichomonas tenax* and periodontal diseases: a concise review. *Parasitology* 144, 1417–1425. doi: 10.1017/S0031182017000701

Matthew, M. A., Yang, N., Ketzis, J., Mukaratirwa, S., and Yao, C. (2023). *Trichomonas tenax*: a neglected protozoan infection in the oral cavities of humans and dogs—a scoping review. *Trop. Med. Infect. Dis.* 8:60. doi: 10.3390/tropicalmed8010060

Mercer, F., and Johnson, P. J. (2018). *Trichomonas vaginalis*: pathogenesis, symbiont interactions, and host cell immune responses. *Trends Parasitol.* 34, 683–693. doi: 10.1016/j.pt.2018.05.006

Molgora, B. M., Mukherjee, S. K., Baumel-Alterzon, S., Santiago, F. M., Muratore, K. A., Sisk Jr, A. E., et al. (2023). *Trichomonas vaginalis* adherence phenotypes and extracellular vesicles impact parasite survival in a novel in vivo model of pathogenesis. *PLoS Negl. Trop. Dis.* 17, e0011693. doi: 10.1371/journal.pntd.0011693

Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N., and Hirt, R. P. (2012). A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. *PLoS ONE* 7:e30287. doi: 10.1371/journal.pone.0030287

Nishimura, D. (2000). RepeatMasker. *Biotech. Softw. Internet Rep.* 1, 36–39. doi: 10.1089/152791600319259

Noël, C. J., Diaz, N., Sicheritz-Ponten, T., Safarikova, L., Tachezy, J., Tang, P., et al. (2010). *Trichomonas vaginalis* vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics. *BMC Genomics* 11:99. doi: 10.1186/1471-2164-11-99

Ong, S. C., Luo, H. W., Cheng, W. H., Ku, F. M., Tsai, C. Y., Huang, P. J., et al. (2024). The core exosome proteome of *Trichomonas vaginalis*. *J. Microbiol. Immunol. Infect.* 57, 246–256. doi: 10.1016/j.jmii.2024.02.003

Petersen, T. N., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Rai, A. K., and Johnson, P. J. (2019). *Trichomonas vaginalis* extracellular vesicles are internalized by host cells using proteoglycans and caveolin-dependent endocytosis. *Proc. Nat. Acad. Sci.* 116, 21354–21360. doi: 10.1073/pnas.1912356116

Ribeiro, L. C., Santos, C., and Benchimol, M. (2015). Is *Trichomonas tenax* a parasite or a commensal? *Protist* 166, 196–210. doi: 10.1016/j.protis.2015.02.002

Riestra, A. M., de Miguel, N., Dessi, D., Simoes-Barbosa, A., and Mercer, F. K. (2022). "*Trichomonas vaginalis*: lifestyle, cellular biology, and molecular mechanisms of pathogenesis," in *Lifecycles of Pathogenic Protists in Humans* (Cham: Springer International Publishing), 541-617.

Salas, N., Pedreros, M. B., dos Santos Melo, T., Maguire, V. G., Sha, J., Wohlschlegel, J. A., et al. (2023). Role of cytoneme structures and extracellular vesicles in *Trichomonas vaginalis* parasite-parasite communication. *Elife* 12:e86067. doi: 10.7554/eLife.86067.sa2

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., and Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. *Genomics* 59, 24–31. doi: 10.1006/geno.1999.5854

Santi-Rocca, J. (2020). "The Protozoome of the Periodontal Sulcus: from health to disease," in *Eukaryome Impact on Human Intestine Homeostasis and Mucosal Immunology: Overview of the First Eukaryome Congress at Institut Pasteur*. (Paris: Springer International Publishing), 113–131.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, A. (2010). Virulence mechanisms of *Tannerella forsythia*. *Periodontol.* 54:106. doi: 10.1111/j.1600-0757.2009.00332.x

Smit, A., and Hubley, R. (2008). *RepeatModeler Open-1.0. Repeat Masker*. Available online at: http://www.repeatmasker.org

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32(Suppl_2), W309-W312. doi: 10.1093/nar/gkh379

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., and Wang, Y. (2023). OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res.* 51, W397-W403. doi: 10.1093/nar/gkad313

Törönen, P., Medlar, A., and Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* 46, W84–W88. doi: 10.1093/nar/gky350

Twu, O., de Miguel, N., Lustig, G., Stevens, G. C., Vashisht, A. A., Wohlschlegel, J. A., et al. (2013). *Trichomonas vaginalis* exosomes deliver cargo to host cells and mediate host: parasite interactions. *PLoS Pathog.* 9:e1003482. doi: 10.1371/journal.ppat.1003482

Viborg, A. H., Terrapon, N., Lombard, V., Michel, G., Czjzek, M., Henrissat, B., et al. (2019). A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16). *J. Biol. Chem.* 294, 15973–15986. doi: 10.1074/jbc.RA119.010619

Wardman, J. F., Bains, R. K., Rahfeld, P., and Withers, S. G. (2022). Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nat. Rev. Microbiol.* 20, 542–556. doi: 10.1038/s41579-022-00712-1

Yang, N., Christie, J., Keen, H. L., Matthew, M. A., and Yao, C. (2022). Draft genome sequence of *Trichomonas tenax* strain Hs-4: NIH. *Microbiol. Res. Announcem.* 11, e00157–e00122. doi: 10.1128/mra.00157-22

Zubáčová, Z., Cimburek, Z., and Tachezy, J. (2008). Comparative analysis of trichomonad genome sizes and karyotypes. *Mol. Biochem. Parasitol.* 161, 49–54. doi: 10.1016/j.molbiopara.2008.06.004

Frontiers in Microbiology

Explores the habitable world and the potential of microbial life

The largest and most cited microbiology journal which advances our understanding of the role microbes play in addressing global challenges such as healthcare, food security, and climate change.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact