frontiers | Research Topics

Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution

Edited by
Paulo Jorge Dias, Satyanarayana Tulasi and Digvijay Verma

Published in
Frontiers in Microbiology
Frontiers in Genetics

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-5650-4
DOI 10.3389/978-2-8325-5650-4

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

Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution

Topic editors

Paulo Jorge Dias — University of Lisbon, Portugal

Satyanarayana Tulasi — University of Delhi, India

Digvijay Verma — Babasaheb Bhimrao Ambedkar University, India

Citation

Dias, P. J., Tulasi, S., Verma, D., eds. (2024). *Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5650-4

Table of contents

05 Editorial: Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution
Digvijay Verma, Tulasi Satyanarayana and Paulo Jorge Dias

08 Potential mobile units drive the horizontal transfer of phytoplasma effector *phyllogen* genes
Ryosuke Tokuda, Nozomu Iwabuchi, Yugo Kitazawa, Takamichi Nijo, Masato Suzuki, Kensaku Maejima, Kenro Oshima, Shigetou Namba and Yasuyuki Yamaji

21 Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing
Wenxuan Zhao, Wei Zeng, Bo Pang, Ming Luo, Yao Peng, Jialiang Xu, Biao Kan, Zhenpeng Li and Xin Lu

30 Genomic characteristics of *Mycobacterium tuberculosis* isolates of cutaneous tuberculosis
You-Ming Mei, Wen-Yue Zhang, Ji-Ya Sun, Hai-Qin Jiang, Ying Shi, Jing-Shu Xiong, Le Wang, Yan-Qing Chen, Si-Yu Long, Chun Pan, Tao Luo and Hong-Sheng Wang

38 Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy
Lezheng Yu, Yonglin Zhang, Li Xue, Fengjuan Liu, Runyu Jing and Jiesi Luo

52 Phylogenomic analysis of the *Porphyromonas gingivalis* - *Porphyromonas gulae* duo: approaches to the origin of periodontitis
Mauricio Morales-Olavarría, Josefa Nuñez-Belmar, Dámariz González, Emiliano Vicencio, Jaime Andres Rivas-Pardo, Cristian Cortez and Juan P. Cárdenas

74 High prevalence of *Helicobacter pylori* mixed infections identified by multilocus sequence typing in Ningbo, China
Yongxiong Zhang, Haiying Gu, Zhouhong Shi, Weiqin Chen, Airu Li, Weiwei Ye, Cheng Zhang, Huikun Yuan and Mingming Zhao

87 Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of *Neisseria meningitidis* leveraging Oxford Nanopore long reads
Zuyu Yang, Andrea Guarracino, Patrick J. Biggs, Michael A. Black, Nuzla Ismail, Jana Renee Wold, Tony R. Merriman, Pjotr Prins, Erik Garrison and Joep de Ligt

100 Emergence of *Raoultella ornithinolytica* in human infections from different hospitals in Ecuador with *OXA*-48-producing resistance
José E. Villacís, Hugo G. Castelán-Sánchez, Jorge Rojas-Vargas, Ulises E. Rodríguez-Cruz, Viviana Albán, Jorge A. Reyes, Pablo M. Meza-Rodríguez, Sonia Dávila-Ramos, Fernando Villavicencio, Margarita Galarza and Monica C. Gestal

113 AI-driven pan-proteome analyses reveal insights into the biohydrometallurgical properties of *Acidithiobacillia*
Liangzhi Li, Lei Zhou, Chengying Jiang, Zhenghua Liu, Delong Meng, Feng Luo, Qiang He and Huaqun Yin

128 The man, the plant, and the insect: shooting host specificity determinants in *Serratia marcescens* pangenome
Anton E. Shikov, Anastasiya V. Merkushova, Iuliia A. Savina, Anton A. Nizhnikov and Kirill S. Antonets

143 Insights into early evolutionary adaptations of the *Akkermansia* genus to the vertebrate gut
Dámariz González, Mauricio Morales-Olavarria, Boris Vidal-Veuthey and Juan P. Cárdenas

160 Genomic diversity and taxonomic marker for *Arcobacter* species
Guilan Zhou, Yixin Gu, Hairui Wang, Xiaoli Chen, Xin Zhang, Zhujun Shao, Xiaomei Yan, Jianzhong Zhang and Maojun Zhang

174 Phylogenetic distribution of malonate semialdehyde decarboxylase (MSAD) genes among strains within the genus *Mycobacterium*: evidence of MSAD gene loss in the evolution of pathogenic mycobacteria
Duhyung Lee, Dong Hyun Kim, Hyejun Seo, Seaone Choi and Bum-Joon Kim

188 Complete genome sequencing and comparative genomic analysis of three donkey *Streptococcus equi* subsp. *equi* isolates
Yuwei Zhang, FenFen Lv, Yan Su, Huan Zhang and Baojiang Zhang

202 RdJ detection tests to identify a unique MRSA clone of ST105-SCC*mec*II lineage and its variants disseminated in the metropolitan region of Rio de Janeiro
Matheus Assis Côrtes Esteves, Alice Slotfeldt Viana, Gabriela Nogueira Viçosa, Ana Maria Nunes Botelho, Ahmed M. Moustafa, Felipe Raposo Passos Mansoldo, Adriana Lucia Pires Ferreira, Alane Beatriz Vermelho, Bernadete Teixeira Ferreira-Carvalho, Paul Joseph Planet and Agnes Marie Sá Figueiredo

210 Differential microbial composition in parasitic vs. questing ticks based on 16S next-generation sequencing
Lin Zhang, Jing Han, Qi Zhou, Zhen He, Shi-Wei Sun, Rui Li, Rui-Shan Li, Wen-Kai Zhang, Yu-Hua Wang, Lin-Li Xu, Zhen-Hua Lu and Zhong-Jun Shao

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 24 October 2024
DOI 10.3389/fgene.2024.1490645

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Samuel A. Cushman,
United States Department of Agriculture,
United States

*CORRESPONDENCE
Digvijay Verma,
digvijay.udsc@gmail.com
Tulasi Satyanarayana,
tsnarayana@gmail.com
Paulo Jorge Dias,
pjdias@tecnico.ulisboa.pt

RECEIVED 03 September 2024
ACCEPTED 10 October 2024
PUBLISHED 24 October 2024

CITATION
Verma D, Satyanarayana T and Dias PJ (2024) Editorial: Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution. *Front. Genet.* 15:1490645. doi: 10.3389/fgene.2024.1490645

COPYRIGHT
© 2024 Verma, Satyanarayana and Dias. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution

Digvijay Verma[1]*, Tulasi Satyanarayana[2]* and Paulo Jorge Dias[3,4]*

[1]Department of Environmental Microbiology, School of Earth and Environmental Sciences, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India, [2]Department of Biological Sciences and Engineering, Netaji Subhas University of Technology, New Delhi, India, [3]iBB - Institute for Bioengineering and Biosciences, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, [4]Associate Laboratory i4HB - Institute for Health and Bioeconomy at Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

KEYWORDS

pangenomics, microbial genomics, next-generation sequencing, whole genome sequencing, metagenome

Editorial on the Research Topic
Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution

Developments in Genomics and DNA Sequencing technologies complement each other. The genesis of Comparative Genomics evolved during the late 70s and early 80s, when the first generation of efficient DNA sequencing technologies had been developed by Walter Gilbert and Frederick Sanger (Shendure et al., 2017). Subsequently, more reliable, and automated sequencing machines were developed by Applied Biosystems and their dissemination to research laboratories in the 90s that led to a major leap in Comparative Genomics (Shendure et al., 2017). Thus, the focus of researchers gradually shifted from small-scale DNA analysis (e.g., genes) to whole genome sequence analysis, an era inaugurated with the release of first complete genome of a free-living organism, the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995), followed by the first sequenced genome of a eukaryotic organism, *Saccharomyces cerevisiae* (Goffeau et al., 1996), culminating in the first draft genome sequence of *Homo sapiens* (Lander et al., 2001; Venter et al., 2001).

The emergence of Next-Generation Sequencing (NGS) technologies was the game changer during the first decade of the 21st century that led to the development of a new sub-field of Genomics, i.e., Pangenomics. The initial concept of Pangenome fulfilled the demand for minimum inclusion of the required number of genomes to achieve good confidence by covering the entire genes of a species (Tettelin et al., 2005). The entire spectrum of genes, gene order, gene content, and their structural variations could be explored to assess the species' phylogenetic relatedness. This gene-oriented branch of genomics categorizes genes into three groups based on their relatedness among various strains of a species. This includes core (shared among 99% of the strains), shell (10–99%), and, cloud (<10% of the strains) genes. The core genes are evolutionarily conserved, while shell and cloud genes

constitute variable regions. Microbial Pangenomics grabbed attention in 2005 when researchers compared whole genome sequences of eight strains of *Streptococcus agalactiae* that unveiled several new genes into the assembly (Tettelin et al., 2005). Palsson's team explored pangenomic analysis of several strains of *Escherichia coli* (Monk et al., 2013) and *Mycobacterium tuberculosis* (Kavvas et al., 2018) and their comparison to identify closely as well as distantly related genes. The second decade of this century witnessed the birth of the 3rd generation of sequencing technologies. This led to the evolution of the concept of Pangenomics. With this new concept, Pangenomics is now viewed as an advanced approach to Genomics that considers the entire genomic diversity of a species or organisms, where every base matters. This also gave rise to a new discipline of Computational Pangenomics, responsible for the development of efficient data structures, algorithms, and statistical methods to perform bioinformatic analyses on computational objects known as Pangenome Graphs.

This Research Topic highlights the duality that exists between Genomics and DNA Sequencing technologies. Many of the sixteen articles presented here used sequencing approaches to explore various facets of Pangenomics and Comparative Genomics. The following text provides a summary of exciting studies from the experts in this area.

Whole genome sequencing (WGS) analysis performed by Zhang et al. on three different isolates of *Streptococcus equi* originating from donkeys showed a high degree of resemblance to each other. A comparative analysis of these strains with a genome sequence of a horse isolate *S. equi* 4047 unveiled several rearrangements and inversions in their genomes. Prophage and other virulence factors were identified to determine their genomic diversity and virulence factors. A similar finding was recorded by Shikov et al., where the group observed host specificity among the members of *Serratia marcescens*. Prophages were specific for humans, while genetic islands were specific for plants and insects. A comparative genomics of Malonate Semialdehyde Decarboxylases (*MSAD*) gene from *Mycobacterium* spp. performed by Lee et al. revealed the absence of MSAD-like genes (*MSAD-1* and *MSAD-2*) in the pathogenic species of *Mycobacterium*. Thus, MSAD loss suggested host-specific adaptations among pathogenic mycobacterial species. In another investigation on *Mycobacterium* genomes, Mei et al. report the isolation of *Mycobacterium tuberculosis* (MTB) strains from patients with Cutaneous Tuberculosis (CTB) and the corresponding genomic characterization. The study concludes that these isolates predominantly belong to the Beijing lineage, consistent with the present epidemic status of the MTB disease in China. Besides, different infection sites of MTB in patients are independent of any association with specific genomic changes. *M. tuberculosis* pangenome was also constructed revealing the lack of significant differences in the accessory genome among different types of strains associated with the CTB disease.

Comparative genomics offers to unveil the alteration in the genomes that occurs due to various adaptations. A comparative pangenome analysis of 371 different genomes of *Arcobactor* species conducted by Zhou et al. successfully identified a taxonomic marker gene (*gene* 711) for effective classification of *Arcobactor* spp. over the ANI (Average Nucleotide Identity) and isDDH (*in silico* DNA–DNA hybridization) like traditional methods of bacterial identification. Yang et al. present a new bioinformatics pipeline based on the PanGenome Graph Builder (PGGB), allowing the semi-automated construction of a Pangenome Graph. The resulting computational object represents the genomic relationships among the highly recombinant genome sequences of *Neisseria meningitidis* strains, showing enhanced detection capability of genetic variations. This pipeline allows the identification of new virulence and antimicrobial resistance genes to facilitate vaccine design and quick detection of bacterial outbreaks, with an impact on public health surveillance. The pangenomics-based variant analysis, population genetics, and evolutionary biology can be studied effectively. Pangenome analysis of *Akkermansia* (the only known genus of Verrucomicrobiota) performed by Dámariz González et al. revealed that the variation in its genome is due to horizontal gene transfer either from unknown species or from Gram-negative gut bacteria. Moreover, the analysis revealed that genes involved in mucin degradation, surface interaction, and, adhesion are constituents of the last *Akkermansia* common ancestor (LAkkCA). The mucin degradation ability of *Akkermansia* is an essential feature of being a symbiotic species. The study conducted by Morales-Olavarría et al. on pangenomics concluded that *Porphyromonas gingivalis* and *Porphyromonas gulae* maintain a strong core-to-accessory ratio for housekeeping genes, and the differences in their genomes are chiefly due to involvement of genes of unknown functions. Esteves et al. developed a binary matrix (0, 1) based protocol to develop effective biomarkers for identifying amorphic sequence mutations. The technique can be employed on other MRSA clones or from other bacterial species. This technique overcomes the problems associated with the WGS and the requirement of trained personnel. Confirming the increasing importance of Artificial Intelligence (AI) algorithms in Comparative Genomics, two studies published in this topic have used deep learning (DL) algorithms to analyze two distinct types of biological problems, the prediction of protein structure and RNA modification. Besides pangenomics, Li et al. developed a proteome analysis using a deep learning (DL) algorithm to study the *Acidithiobacillus* genome (a well-known bacterium involved in bioleaching process for metal recovery from ores). DL assisted in gene ontology in terms of 93.6% unknown proteins. To date, the crystal structures of only 14 distinct proteins of this bacterium are known, further studies are welcome in this line of work. Yu et al. have taken advantage of autoBioSeqpy, a Keras-based deep learning software for fast and easy development, training, and analysis of deep learning model architectures for biological sequence classification, to develop a DL model able to predict the 5-methyluridine (m5U) modification. The authors concluded that the model combining convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) was consistently better than all other DL models tested, being named Deepm5U. This model showed a predictive performance that outperformed the current state-of-the-art tool.

Zhao et al. use long-read sequencing based on the latest generation of Oxford Nanopore Technology (ONT) to identify MultiDrug Resistance (MDR) genes in *Salmonella* strains. The authors focus their attention on the spread of MDR-associated genes carried out by plasmids since these mobile genetic elements can be transferred easily between different bacterial

species, enabling a quick and effective dissemination of such genes, and causing important public health issues. WGS was used by Villacís et al. in the identification of the pathogen *Raoultella ornithinolytica*. These authors propose the replacement of the automated identification methodology based on phenotypic characteristics with this new methodology to ensure accuracy. This study also reports an evolutionary origin of certain MDR genes that are relatively recent in this human pathogen, a fact consistent with the open Pangenome attributed to *R. ornithinolytica* and with the worrying trend of quick acquisition and dissemination of antibiotic resistance genes observed in this *Enterobacteriaceae* species. WGS was also explored by Tokuda et al. to analyze phyllody-inducing genes (also known as phyllogens) in various *Phytoplassma* species and strains. The authors showed that the flanking sequences of these genes act as Potential Mobile Units (PMUs), facilitating the horizontal transfer of phyllogens. This study also showed that phyllogen function and the corresponding coding sequences are highly conserved, a fact suggesting that these genes are essential for survival of their *Phytoplasma* hosts. Finally, in this Research Topic, two research groups have analyzed important human pathogens and the corresponding epidemic status in different regions of China. In the first study, Zhang et al. used NGS to determine the sequence of 16S rDNA gene of the whole microbiome of parasitic ticks in Wuwei City, showing the existence of a wide range of bacterial species and high phylogenetic diversity in the corresponding microbiomes. A network analysis of the detected bacterial genera showed that depending on the genera, microbial combinations might be promotive or inhibitive of co-infection. In another study, Zhang et al. analyzed the most common human pathogen responsible for gastrointestinal diseases, *Helicobacter pylori*, in Ningbo, China. These authors used multi-locus sequence typing to show that this bacterial species has a high prevalence of mixed infections, having identified 246 new alleles and 53 new sequence types. This study also confirmed a high genomic diversity present in this human pathogen in this country and showed the prevalence of mixed infections in Ningbo was higher than the ones reported in other regions of China.

Looking ahead, the awesome ability of Artificial Intelligence in the recognition of patterns, and the anticipated dissemination of these methodologies to all areas of scientific knowledge combined with the advent of the first commercial generation of quantum computers at the turn of the decade and the corresponding leap in the computational power available to biologists anticipate a new major leap in Comparative Genomics and Pangenomics. The development of new methodologies combining these technological revolutions suggests that soon biologists will be able to directly simulate the phenotype of microbial strains and species from the observed genotypes, leading to the evolution of the current descriptive nature of Comparative Genomics and Pangenomics towards a new one more predictive. The future thus looks bright.

Author contributions

DV: Writing–original draft, Writing–review and editing. TS: Writing–original draft, Writing–review and editing. PD: 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.

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.

References

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (5223), 496–512. doi:10.1126/science.7542800

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science.* 274(5287):546, 563–567. doi:10.1126/science.274.5287.546

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., et al. (2018). Machine learning and structural analysis of *Mycobacterium tuberculosis* pangenome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* 9, 4306. doi:10.1038/s41467-018-06634-y

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., et al. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U. S. A.* 110 (50), 20338–20343. doi:10.1073/pnas.1307797110

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present, and future. *Nature* 550 (7676), 345–353. doi:10.1038/nature24286

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102 (39), 13950–13955. doi:10.1073/pnas.0506758102

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291 (5507), 1304–1351. doi:10.1126/science.1058040

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 11 May 2023
DOI 10.3389/fgene.2023.1132432

Check for updates

OPEN ACCESS

EDITED BY
Satyanarayana Tulasi,
University of Delhi, India

REVIEWED BY
Inderjit Singh Yadav,
Punjab Agricultural University, India
Jacqueline Fletcher,
Oklahoma State University, United States
Valeria Trivellone,
University of Illinois at Urbana-Champaign, United States

*CORRESPONDENCE
Kensaku Maejima,
amaejima@g.ecc.u-tokyo.ac.jp

RECEIVED 27 December 2022
ACCEPTED 03 April 2023
PUBLISHED 11 May 2023

CITATION
Tokuda R, Iwabuchi N, Kitazawa Y, Nijo T, Suzuki M, Maejima K, Oshima K, Namba S and Yamaji Y (2023), Potential mobile units drive the horizontal transfer of phytoplasma effector *phyllogen* genes. *Front. Genet.* 14:1132432. doi: 10.3389/fgene.2023.1132432

COPYRIGHT
© 2023 Tokuda, Iwabuchi, Kitazawa, Nijo, Suzuki, Maejima, Oshima, Namba and Yamaji. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Potential mobile units drive the horizontal transfer of phytoplasma effector *phyllogen* genes

Ryosuke Tokuda[1], Nozomu Iwabuchi[1], Yugo Kitazawa[1], Takamichi Nijo[1], Masato Suzuki[1], Kensaku Maejima[1]*, Kenro Oshima[2], Shigetou Namba[1] and Yasuyuki Yamaji[1]

[1]Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan, [2]Faculty of Bioscience and Applied Chemistry, Hosei University, Tokyo, Japan

Phytoplasmas are obligate intracellular plant pathogenic bacteria that can induce phyllody, which is a type of abnormal floral organ development. Phytoplasmas possess phyllogens, which are effector proteins that cause phyllody in plants. Phylogenetic comparisons of *phyllogen* and 16S rRNA genes have suggested that *phyllogen* genes undergo horizontal transfer between phytoplasma species and strains. However, the mechanisms and evolutionary implications of this horizontal gene transfer are unclear. Here, we analyzed synteny in *phyllogen* flanking genomic regions from 17 phytoplasma strains that were related to six '*Candidatus*' species, including three strains newly sequenced in this study. Many of the *phyllogens* were flanked by multicopy genes within potential mobile units (PMUs), which are putative transposable elements found in phytoplasmas. The multicopy genes exhibited two distinct patterns of synteny that correlated with the linked *phyllogens*. The low level of sequence identities and partial truncations found among these *phyllogen* flanking genes indicate that the PMU sequences are deteriorating, whereas the highly conserved sequences and functions (e.g., inducing phyllody) of the *phyllogens* suggest that the latter are important for phytoplasma fitness. Furthermore, although their *phyllogens* were similar, PMUs in strains related to '*Ca*. P. asteris' were often located in different regions of the genome. These findings strongly indicate that PMUs drive the horizontal transfer of *phyllogens* among phytoplasma species and strains. These insights improve our understanding of how symptom-determinant genes have been shared among phytoplasmas.

KEYWORDS

phyllogen, phytoplasma, effector, horizontal gene transfer, potential mobile unit, pathogenicity, symptom

Introduction

Gene acquisition by horizontal gene transfer (HGT) plays a crucial role in the adaptive evolution of most organisms (Keeling and Palmer, 2008; Soucy et al., 2015). In bacteria, HGT is generally mediated by transposable elements (e.g., phages, plasmids, and transposons) that can travel between host bacteria, as well as by homologous recombination between genomes (Soucy et al., 2015; Drew et al., 2021). Horizontal transfer of single genes or genomic islands

can influence bacterial phenotypes and pathogenicity (Melnyk et al., 2019; Gluck-Thaler et al., 2020; Drew et al., 2021). In the genus *Rhizobium*, there is a plasmid carrying genes that can influence root nodulation transfer among different species and determines the host range for root nodule induction (Hooykaas et al., 1982). In the family *Xanthomonadaceae*, colonization of plant vascular tissues prior to systemic infection was associated with horizontal transfer of a single gene, *cbsA* (Gluck-Thaler et al., 2020). The mode of HGT-mediated acquisition of pathogenic genes may have profound implications for pathogenic microbial evolution. These HGTs are mainly inferred using parametric or phylogenetic methods (Ravenhall et al., 2015).

Phytoplasmas ['*Candidatus* (*Ca.*) Phytoplasma' spp.] are obligate intracellular plant pathogenic bacteria in the class Mollicutes. Phytoplasmas are transmitted by phloem-feeding insects of the order Hemiptera and can infect more than 1,000 plant species (Marcone, 2014). Although phytoplasma genomes are small and lack many metabolic genes (Oshima et al., 2004), multicopy genes account for 18%–28% of all phytoplasma genes (Oshima et al., 2004; Bai et al., 2006; Kube et al., 2008; Tran-Nguyen et al., 2008). Most of these multicopy genes occur in gene clusters called potential mobile units (PMUs; Bai et al., 2006; Arashida et al., 2008a; Wei et al., 2008). PMUs are often associated with *tra5*, which is a putative transposable gene belonging to the IS3 family, among other genes such as *fliA*, *ssb*, *dam*, *himA*, *hflB*, *smc*, *tmk*, *dnaB*, and *dnaG* (Arashida et al., 2008a). Some PMUs have been reported to transfer in phytoplasma genomes and to exist as probably transposable circular extrachromosomal elements (Arashida et al., 2008a; Toruño et al., 2010; Ku et al., 2013). In addition, PMUs can reportedly undergo HGT between phytoplasma genomes, which mediates the acquisition of novel genes by phytoplasmas (Chung et al., 2013; Ku et al., 2013; Music et al., 2019). Although many PMUs have lost their capacity for HGT due to the deletion of genes such as *tra5*, they are still described as PMUs in many studies, including this one.

Phytoplasma diseases are associated with unique symptoms such as dwarfing, witches' broom, yellowing, and phyllody (Namba et al., 1993). These symptoms are mainly induced by effector proteins secreted by the phytoplasmas (Hoshi et al., 2009; MacLean et al., 2011; Sugio et al., 2011; Minato et al., 2014; Huang et al., 2021). Phytoplasmas have effectors that cause different disease symptoms. Many of the secreted effector proteins are encoded by PMUs, which phytoplasmas may have acquired by HGT (Sugio and Hogenhout, 2012; Ku et al., 2013). However, to understand the genetics of phytoplasma pathogenicity, it is necessary to confirm that the acquisition of these effectors is associated with HGT of PMUs.

Phyllogens are effector family proteins unique to phytoplasmas that induce phyllody in various eudicots (MacLean et al., 2011; Maejima et al., 2014; Yang et al., 2015; Kitazawa et al., 2017; Iwabuchi et al., 2020). The conserved molecular mechanisms responsible for inducing phyllody have been elucidated (MacLean et al., 2014; Maejima et al., 2014; Kitazawa et al., 2022). Phyllogen genes (*phyllogens*) can be phylogenetically divided into four groups (i.e., phyl-A, -B, -C, and -D) and the evolutionary history of these *phyllogens* differs from that of phytoplasmas, as confirmed by analyses of phytoplasma 16S rRNA gene sequences (Iwabuchi et al., 2020; Kalla et al., 2021). Indeed, different groups of *phyllogens* have been found in the same '*Ca.* Phytoplasma' sp. These findings suggest that *phyllogens* have undergone HGT among various phytoplasma species, thereby enabling these phytoplasmas to induce phyllody (Iwabuchi et al., 2020).

However, gene sequence information is insufficient to establish the evolutionary history of all genes and reveal the mechanism of HGT; therefore, a better understanding of the structure of the gene flanking regions may be necessary (Melnyk et al., 2019; Gluck-Thaler et al., 2020). *Phyllogens* are often found within PMUs (Jomantiene et al., 2007; Sugio and Hogenhout, 2012; Maejima et al., 2014; Wang et al., 2018; Huang et al., 2022); therefore, HGT of *phyllogens* may be mediated by PMUs (Iwabuchi et al., 2020). However, *phyllogen* flanking PMUs have only been analyzed in a few phytoplasma strains (Sugio and Hogenhout, 2012; Wang et al., 2018; Huang et al., 2022), and most *phyllogen* flanking regions have not been studied in detail.

In this study, we analyzed the sequences of *phyllogen* flanking regions from various phytoplasma species and strains, and evaluated similarities to infer the mechanisms and evolutionary history of *phyllogen* acquisition. Sequencing and comparison of *phyllogen* flanking regions revealed synteny among PMU-associated genes for each group of *phyllogens*, indicating that horizontal transfer of PMUs was associated with *phyllogen* acquisition. Furthermore, HGT events were followed by hypermutation, pseudogenesis, and deletion of PMU-associated genes, whereas the sequences of *phyllogens* were highly conserved.

Materials and methods

Phytoplasma DNA

DNA samples from '*Ca.* P. asteris' HP, RhY, and PaWB-Japan strains, and from '*Ca.* P. phoenicium' PEY were extracted from infected *Hydrangea* spp. (Takinami et al., 2013), *Rhus javanica* (Takinami et al., 2013), *Paulownia tomentosa* (Kakizawa et al., 2006), and periwinkle (Maejima et al., 2014), respectively.

Identification of genomic sequences flanking *phyllogens*

The phytoplasma species and strains used in this study are listed in Supplementary Table S1. Contigs of *phyllogen* sequences were created from draft genomes of the HP, RhY, and PaWB-Japan strains sequenced in this study (see Supplementary Methods), as well as from phytoplasma genomic sequences deposited in GenBank (Supplementary Table S1); we performed tBLASTn searches using amino acid sequences from *phyllogen* homologs (Iwabuchi et al., 2020). To determine upstream sequences for the RhY, PaWB-Japan, and PEY strain contigs, we performed polymerase chain reaction (PCR) using primer pairs that targeted each contig and conserved regions of phytoplasma *fliA*, which is frequently found in the first or second open reading frame (ORF) of PMUs (Arashida et al., 2008a; Wang et al., 2018; Supplementary Table S2). The flanking regions of the RhY and PaWB-Japan strain contigs were determined by PCR using primer pairs designed to match the corresponding regions of the HP and PaWB-China genomes, respectively (Supplementary

Table S2). The downstream region for the PEY strain was determined by PCR using primer pairs that targeted the *phyllogen* sequence and *dnaG* sequence in the AY-WB strain PMU (Supplementary Table S2). PCR was performed using LA Taq (TaKaRa Bio, Inc., Shiga, Japan) and 0.3 μM of each primer, in accordance with the manufacturer's instructions. Amplified fragments (>1.5 kbp) were purified and sequenced using PCR primers, followed by primer walking. Amplified fragments (<1.5 kbp) were purified, cloned into a pCR2.1-TOPO vector (Invitrogen, Carlsbad, CA, United States), and sequenced by Sanger sequencing. DNA samples from phytoplasma strains containing phyl-A group *phyllogens* (Supplementary Table S1) were subjected to PCR amplification of the region between *hflB*-like genes with similarity to *hflB* and a region downstream of *phyllogens* using primers described previously (Jomantiene et al., 2007; Supplementary Table S2). The *phyllogen* flanking sequences were added to the draft genomes and deposited in DDBJ under the accession numbers listed in Supplementary Table S1.

Comparative analysis of *phyllogen* flanking regions

The genomic regions that flanked the *phyllogens* were annotated for protein coding genes using MetaGeneAnnotator (Noguchi et al., 2008) in DFAST (Tanizawa et al., 2018) and analyzed using BLASTp searches restricted to sequences >8,000 bp (Altschul et al., 1990; Camacho et al., 2009). Homologous genes were identified when the E-value, identity, and query-cover between two genes were < 1e−6, >30%, and >75%, respectively. To identify genomic locations of *phyllogen* flanking regions in strains related to '*Ca*. P. asteris,' genomic regions were subjected to comparative analyses with other genomes of strains related to '*Ca*. P. asteris' using BLASTp. The results of these comparative analyses were displayed using Clinker ver. 0.0.21 software to enable gene cluster comparisons (Gilchrist and Chooi., 2021). Pairwise sequence identities were calculated using the Sequence Demarcation Tool (SDT; ver. 1.2) (Muhire et al., 2014). Pseudogenes split into multiple ORFs due to premature stop codons were identified by combining the ORFs.

Alignment and phylogenetic analyses

The nucleotide sequences of *phyllogen*, *fliA*, *himA*, and *hflB* were picked up from the list in Supplementary Table S3 and were aligned using the MUSCLE algorithm in MEGA X software (Kumar et al., 2018). The neighbor-joining (NJ) method (Saitou and Nei, 1987) was used to create phylogenetic trees of these genes using MEGA X software (Kumar et al., 2018). The whole/draft genome-based phylogenetic analysis of phytoplasma strains related to '*Ca*. P. asteris', except for NJAY strain (GenBank accession number MAPF00000000; completeness of only 91%) and TW strain (GenBank accession number QGKT00000000; draft genome reportedly contaminated with two different strains; Cho et al., 2020), was performed as follows. Single-copy genes shared by the '*Ca*. P. asteris'-related phytoplasma genomes (Supplementary Table S3) were identified using SonicParanoid software ver. 1.3.5 (Cosentino and Iwasaki, 2019). The identified homologs were aligned as described above. Aligned sequences were combined for each strain in the same order, and were subjected to phylogenetic analyses using the NJ method as described above.

In planta expression of phyllogen

A modified tobacco rattle virus (TRV)-based gene expression vector system (Iwabuchi et al., 2019) was used to express phyllogen *in planta*. In brief, the $PHYL_{RP166}$ and $PHYL_{NCHU2019}$ sequences were optimized for plant codons and synthesized by ThermoFisher Scientific (Waltham, MA, United States; Supplementary Table S4). The fragments were cloned into pTRV2-2A-sGFP with the primers shown in Supplementary Table S2 by replacing the sGFP region using the NEBuilder HiFi DNA Assembly Cloning kit (New England Biolabs, Ipswich, MA, United States). $PHYL_{HP}$ and $PHYL_{OY}^{E33}$ were cloned with the primers shown in Supplementary Table S2 by adding a single amino acid mutation into pTRV2-cloned $PHYL_{JWB}$ (Iwabuchi et al., 2020) and $PHYL_{OY}$ (Iwabuchi et al., 2019), respectively, using the GeneArt site-directed mutagenesis system (Invitrogen). $PHYL_{PEY}$ was cloned with the primers shown in Supplementary Table S2 by incorporating a two-amino acid mutation into pTRV2-cloned $PHYL_{RP166,}$ amplifying the insert by PCR, and amplifying the rest of the plasmid by inverse PCR. The amplified fragments were ligated using the NEBuilder HiFi DNA Assembly Cloning kit (New England Biolabs). *Arabidopsis thaliana* was maintained as described previously (Iwabuchi et al., 2020). *Agrobacterium tumefaciens* EHA105 cells containing pTRV1 and pTRV2-empty, pTRV2-$PHYL_{OY}$, pTRV2-$PHYL_{RP166}$, pTRV2-$PHYL_{PEY}$, pTRV2-$PHYL_{NCHU2019}$, pTRV2-$PHYL_{HP}$, pTRV2-$PHYL_{JWB,}$ or pTRV2-$PHYL_{OY}^{E33}$ were adjusted to an OD_{600} of 0.1, mixed at a ratio of 1:1, and co-infiltrated into 2- and 3-week-old *A. thaliana* as described previously (Takahashi et al., 2006).

Results

The phylogeny of *phyllogens* was consistent with the synteny of the PMUs

To understand the mechanisms involved in the HGT of *phyllogens* across phytoplasmas, we compared *phyllogen* flanking regions among the genomes of 17 phytoplasma strains related to six species (Figure 1; Supplementary Table S1). Strains harboring the phyl-B group *phyllogens* were not analyzed because the corresponding genomes have not been characterized. Prior to this analysis, we generated draft genome data for the '*Ca*. P. asteris' HP, RhY, and PaWB-Japan strains (Supplementary Figure S1; Supplementary Table S5). The draft genome data were used to generate full-length *phyllogen* sequences for the RhY and PaWB-Japan strains, as well as sequences for *phyllogen* flanking regions from the HP, RhY, and PaWB-Japan strains. The detailed results about genome sequencing were described in Supplementary Results. The *phyllogen* flanking regions for the '*Ca*. P. phoenicium' PEY strains were also determined using PCR. Phylogenetic analyses of *phyllogens* showed that the genes of the 17 strains could be separated into three clades (Figure 1A) corresponding to the phyl-A (6 strains), phyl-C (1 strains), and phyl-D groups (10 strains) described by a previous study (Iwabuchi et al., 2020); the RhY and PaWB-Japan *phyllogens* belong to the phyl-D group. In most *phyllogen* flanking regions (all phyl-A and phyl-C, and

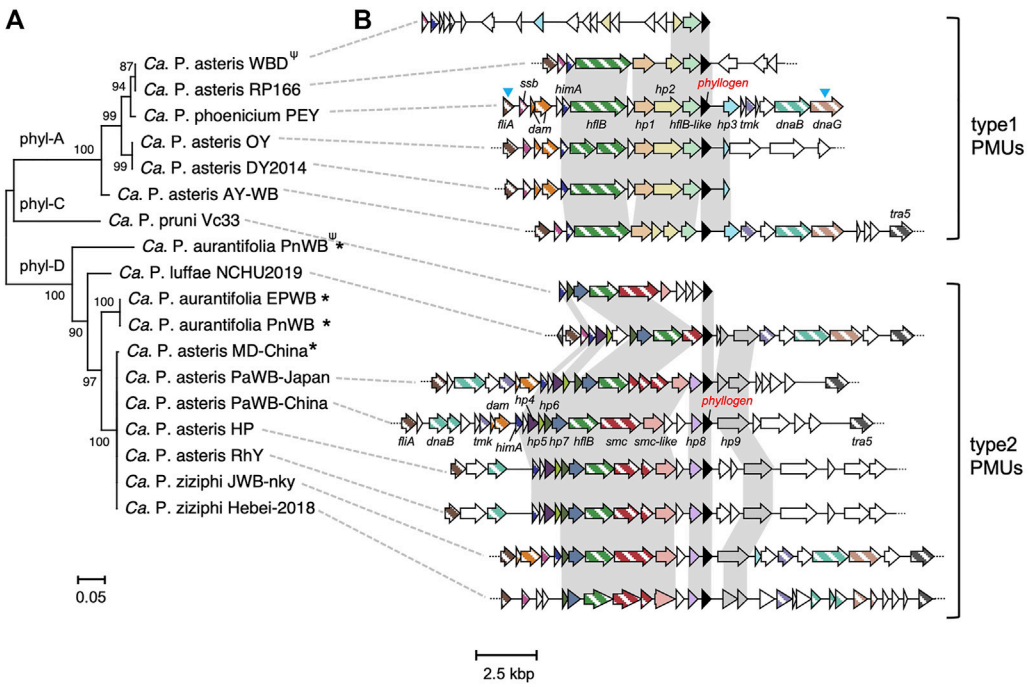


FIGURE 1
Comparative analysis of *phyllogen* flanking regions. **(A)** Neighbor-joining phylogenetic tree for *phyllogens* with known flanking regions of >8,000 bp, analyzed with the complete deletion option. The MUSCLE multiple alignment algorithm (Kumar et al., 2018) was used to align the *phyllogen* nucleotide sequences. Numbers at the nodes represent bootstrap values for 1,000 replicates (only values >70% are shown). Phylogenetic groups (Iwabuchi et al., 2020) are indicated on the tree branches. The scale bar indicates the number of nucleotide substitutions per site. For *phyllogens* truncated at the C-terminus due to premature stop codons (indicated by Ψ), the nucleotide regions after the stop codons were also included in the MUSCLE alignment (Kumar et al., 2018). Phytoplasmas marked with asterisks had *phyllogens* that were not located within PMUs. The '*Candidatus* P. ziziphi' JWB-nky and '*Ca*. P. luffae' NCHU2019 strains had several *phyllogens* with identical sequences; one gene is shown. Full strain names and GenBank accession numbers are listed in Supplementary Table S1. **(B)** Genomic structure of the *phyllogen* flanking regions. Homologous genes present in at least five strains are shown in the same color. *Phyllogens* are shown in black. Diagonal stripe patterns indicate multicopy genes associated with PMUs (Arashida et al., 2008a). The different types of PMU described in this study are indicated on the right. Syntenies unique to the type 1 or type 2 PMUs are highlighted against a light gray background. Open reading frames (ORFs) marked with blue arrowheads have not been completely sequenced. Broken lines at the ends of genomic structures indicate that these regions have been sequenced but are not shown in the figure. The gene names correspond to locus tags in draft genomes (Supplementary Table S8).

seven phyl-D group *phyllogens*), there were several PMU-associated genes (*fliA*, *ssb*, *dam*, *himA*, *hflB*, *smc*, *tmk*, *dnaB*, *dnaG*, and *tra5*; Figure 1B; Arashida et al., 2008a). This observation is consistent with previous reports of phytoplasma strains with *phyllogens* within PMUs (Sugio and Hogenhout, 2012; Maejima et al., 2014).

Except for those from the WBD strain, *phyllogen* flanking PMUs frequently contained genes (e.g., *fliA*, *himA*, and *hflB*) in their upstream regions. These PMUs could be categorized as type 1 or 2 based on some characteristic genes and synteny (Figure 1B). PMU typing was already conducted based on the order of *tmk* and *dnaB* in a previous paper (Huang et al., 2022). Since these genes were not all located in the *phyllogen* flanking PMUs, we classified them based on other genes as described below.

The type 1 PMUs were found in six strains ('*Ca*. P. asteris' AY-WB, DY 2014, OY, RP166, WBD, and '*Ca*. P. phoenicium' PEY), all of which contained phyl-A group *phyllogens*. These type 1 PMUs also contained *hypothetical protein 1* (*hp1*), *hp2*, and an *hflB*-like gene in their upstream regions (Figure 1B). Additionally, the type 1 PMUs exhibited conserved synteny of *fliA*, *ssb*, *himA*, *hflB*, *hp1*, *hp2*, *hflB*-like, *phyllogen*, and *hp3*, except for the PMUs of the WBD strain. In the WBD strain, only the synteny of *hp2*, *hflB*-like, and the *phyllogens* was conserved. In the '*Ca*. P. pruni' CP and '*Ca*. P. trifolii' CPS strains, as well as in some strains related to '*Ca*. P. asteris,' only the region between the *hflB*-like genes and the intergenic regions downstream of the *phyllogens* were characterized; these regions were conserved in all of these strains, and in the OY and WBD strains (Supplementary Figure S2). This result suggests that the type 1 PMUs that harbor phyl-A group *phyllogens* are conserved in '*Ca*. P. pruni,' '*Ca*. P. trifolii,' '*Ca*. P. asteris,' and '*Ca*. P. phoenicium'.

Type 2 PMUs were found in seven strains that have phyl-D group *phyllogens* ('*Ca*. P. asteris' HP, RhY, PaWB-Japan, PaWB-China, '*Ca*. P. luffae' NCHU2019, '*Ca*. P. ziziphi' JWB-nky, and Hebei-2018). Type 2 PMUs were also found in the '*Ca*. P. pruni' Vc33 strain, which has a phyl-C group *phyllogen*. The type 2 PMUs contained *hp4*, *hp5*, *hp6*, *hp7*, *smc*, and *smc*-like genes with similarity to *smc* and *hp8* genes in the upstream region and *hp9* genes in the downstream region. In the PaWB-Japan, PaWB-China, HP, and RhY strains, the type 2 PMUs exhibited conserved synteny of *fliA*, *himA*, *hp4*, *hp5*, *hp6*, *hp7*, *hflB*, *smc*, *smc*-like genes, *hp8*, *phyllogen*, and *hp9*. The JWB-nky and Hebei-2018 strains had another *phyllogen*, and its flanking region also contained genes that were characteristic of type 2 PMUs, although synteny was only partly conserved (Supplementary Figure S3A). The NCHU2019 strain had

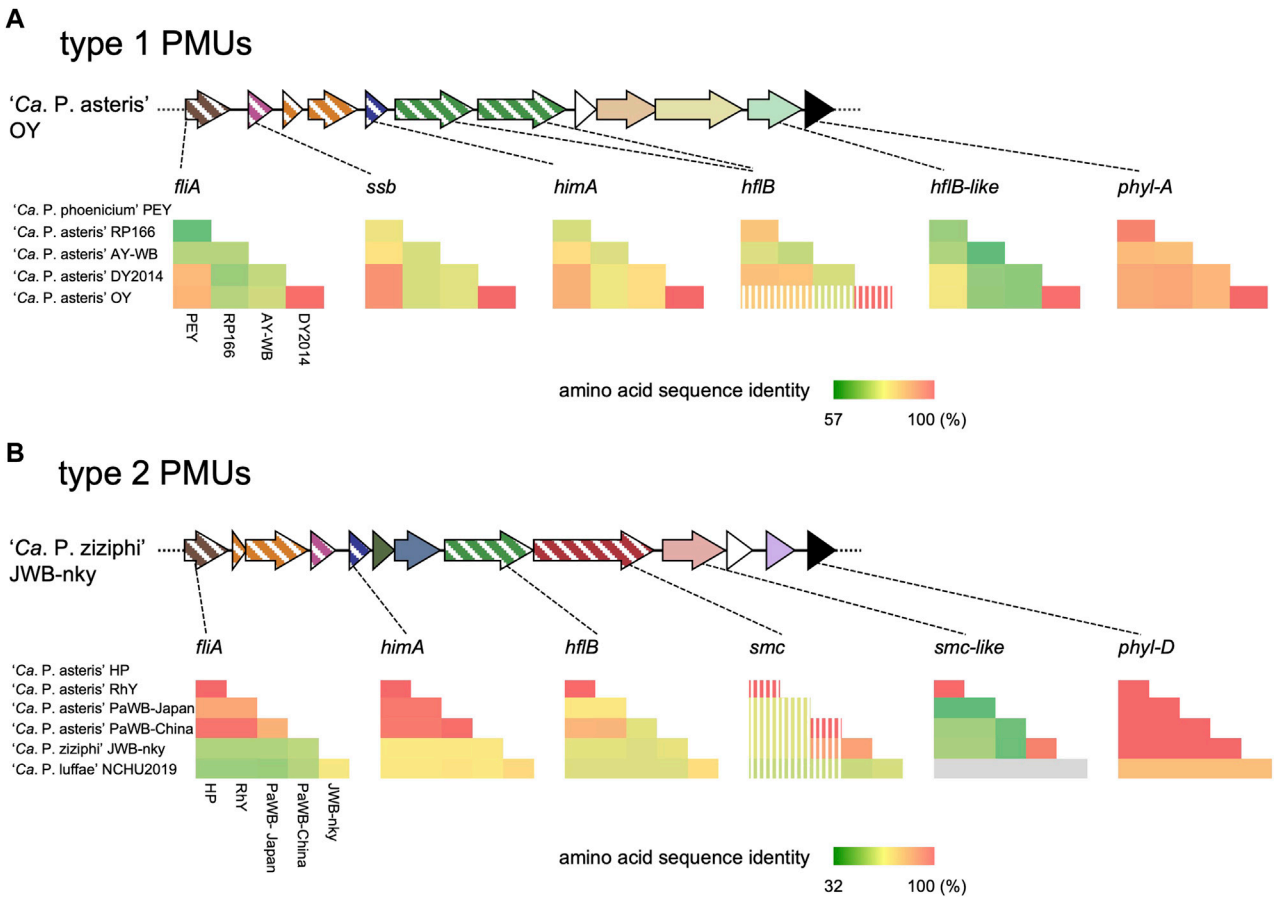


FIGURE 2
Pairwise sequence comparisons of ORFs flanking phyl-A **(A)** and phyl-D **(B)** group *phyllogens*. Genomic structures are represented as in Figure 1B. Pairwise amino acid sequence identities between phytoplasma species/strains are represented by heatmaps. Strain names are shown on the left and below the heatmaps; full names are listed in Supplementary Table S1. Stripes on the heatmap indicate that truncated genes were used in the analysis. The absence of an ORF from the corresponding genomic region of the NCHU2019 strain is indicated in gray. Results are shown for one of two *phyllogens* for '*Ca*. P. ziziphi' and one of three *phyllogens* for '*Ca*. P. luffae.'

three type 2 PMUs with the same gene synteny (Supplementary Figure S3B). These results suggest that type 2 PMUs harboring phyl-D or phyl-C group *phyllogens* are conserved in four different species ('*Ca*. P. asteris,' '*Ca*. P. luffae,' '*Ca*. P. ziziphi,' and '*Ca*. P. pruni'). Additionally, phylogenetic trees based on genes shared by type 1 and 2 PMUs (i.e., *fliA*, *himA*, and *hflB*) indicate that the evolutionary history of these genes differs from that of phytoplasmas. In particular, the *hflB* tree formed two clades, with type 1 and 2 PMU strains (Supplementary Figure S4). Thus, our data indicate that the phylogeny of *phyllogens* is consistent with PMU type (Figure 1).

On the other hand, several *phyllogens* were not located within PMUs (indicated by asterisks; Figure 1). No PMU-associated genes were found in the 20 kbp upstream or downstream of the phyl-D group *phyllogen* in '*Ca*. P. aurantifolia' PnWB, or in the '*Ca*. P. asteris' MD-China strains (except for a *dam* gene in the MD-China strain; Supplementary Figure S6). Additionally, the phyl-D group *phyllogen* in the '*Ca*. P. aurantifolia' EPWB strain was located near the PMU-associated genes, but outside of PMU regions (Supplementary Figure S5).

Phyllogen genes were highly conserved but *phyllogen* flanking PMU genes were not

Some PMUs contain fewer genes than PMUs reported to exist as probably transposable circular extrachromosomal elements such as AY-WB PMU1 (Toruño et al., 2010) and some PMU-associated genes are truncated (Bai et al., 2006; Ku et al., 2013). To understand the evolution of PMUs harboring *phyllogens*, we analyzed the conservation of genes in type 1 and 2 PMUs using pairwise amino acid sequence comparisons. In the type 1 PMUs, *phyllogens* were highly conserved (87%–100%) between '*Ca*. P. asteris' and '*Ca*. P. phoenicium.' However, genes in the *phyllogen* flanking region (*fliA*, *ssb*, *himA*, *hflB*, *hp1*, *hp2*, and *hflB*-like) exhibited low sequence identities (>57%), except in the closely related OY and DY2014 strains (Supplementary Table S6; Figure 2A). In the type 2 PMUs, *phyllogens* were also highly conserved (91%–100%) among '*Ca*. P. asteris', '*Ca*. P. luffae,' and '*Ca*. P. ziziphi,' except for the Vc33 strain *phyllogen* that belongs to a different phyl-C group. However, the genes in the *phyllogen* flanking regions (*fliA*, *himA*, *hp6*, *hp7*, *hflB*, *smc*, *smc*-like, and *hp9*) exhibited

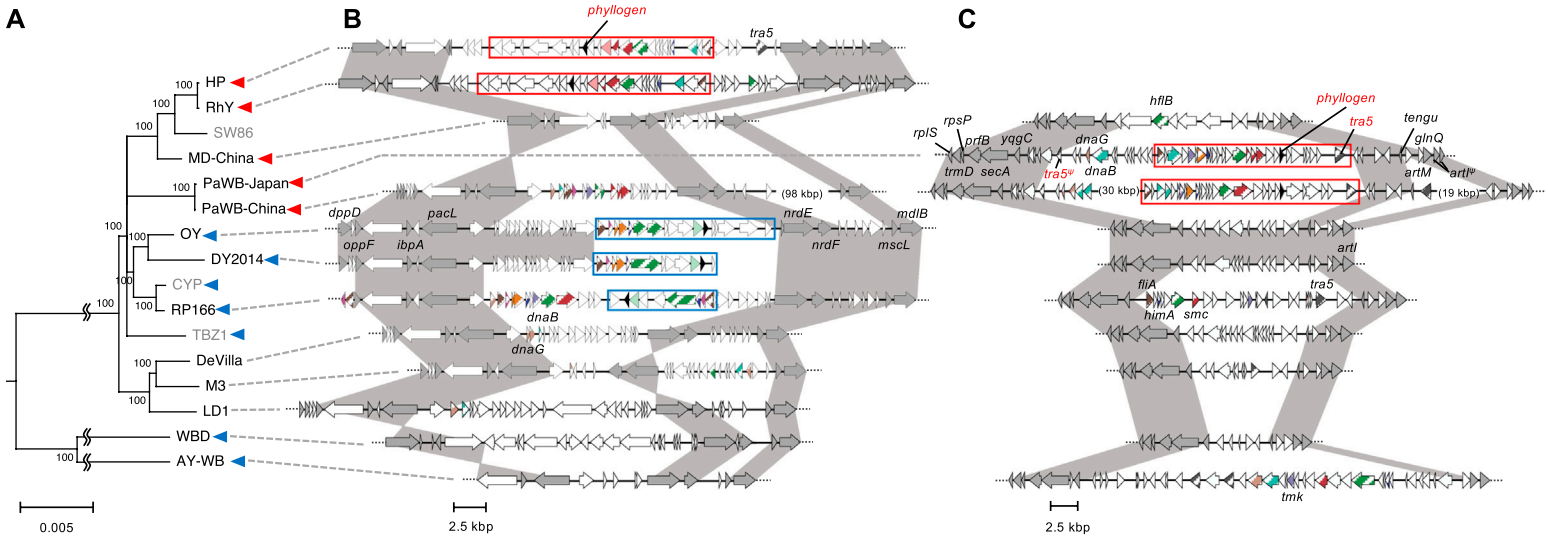


FIGURE 3
Genomic structure of regions flanking PMUs harboring *phyllogens*. **(A)** Whole/draft genome-based phylogenetic tree of '*Ca*. P. asteris' with the distribution of *phyllogens*. The tree was constructed using the neighbor-joining method and includes single-copy genes shared by all the strains related to '*Ca*. P. asteris' listed in Supplementary Table S3. The '*Ca*. P. meliae' ChTYXIII strain was used as an outgroup. Numbers at the nodes represent bootstrap values for 1,000 replicates (only values >70% are shown). The scale bar indicates the number of nucleotide substitutions per site. The full names and GenBank accession numbers of the strains used in the analysis are listed in Supplementary Table S3. The conserved regions shown in **(B)** and **(C)** were not linked in the strains colored gray. Blue and red arrowheads indicate the presence of phyl-A and phyl-D groups in the genome, respectively. **(B, C)** Genomic structure of the regions surrounding PMUs harboring *phyllogens* from HP, RhY, OY, DY 2014, and RP166 strains **(B)**, and PaWB-Japan and PaWB-China strains **(C)**, compared with other '*Ca*. P. asteris' genomes. ORF structures are represented as in Figure 1B. Dark gray ORFs indicate single-copy ORFs annotated by MetaGeneAnnotator (Noguchi et al., 2008). *Phyllogens* are shown in black. Diagonal stripe patterns indicate multicopy genes associated with PMUs (Arashida et al., 2008a). ORFs that are not associated with PMUs are shown in white. Truncated ORFs are marked by Ψ. Conserved regions in strains related to '*Ca*. P. asteris' are shown in dark gray. Type 1 and 2 PMUs that flank *phyllogens* are shown in Figure 1 enclosed by blue and red borders, respectively; the corresponding synteny is highlighted in Figure 1. Broken lines at the ends of genomic structures indicate that these regions have been sequenced but are not shown in the figure.

10.3389/fgene.2023.1132432

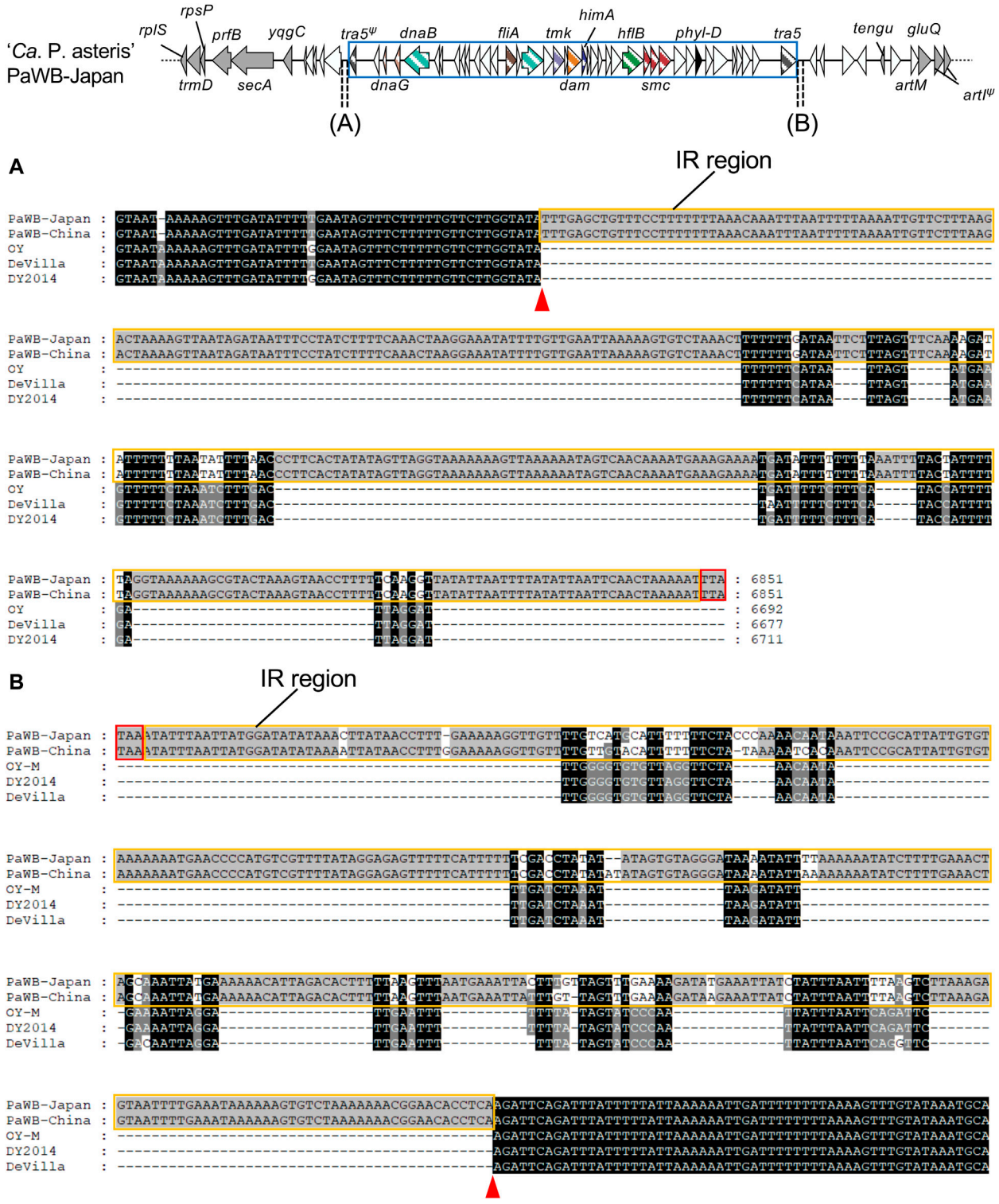


FIGURE 4
Nucleotide sequence alignment of the downstream sequences, beginning at stop codons of truncated *tra5* in upstream regions **(A)** or intact *tra5* in downstream regions **(B)** of *phyllogens*. Stop codons in upstream regions of truncated *tra5* genes were identified by comparing nucleotide sequences with the intact *tra5* gene. *Tra5* stop codons are enclosed by red borders. Inverted repeat regions are enclosed by orange borders. Putative PMU boundaries are marked with red arrowheads. Genomic structures of the *phyllogen* flanking regions in the PaWB-Japan strain are represented as in Figure 3. PMUs are enclosed by blue borders.

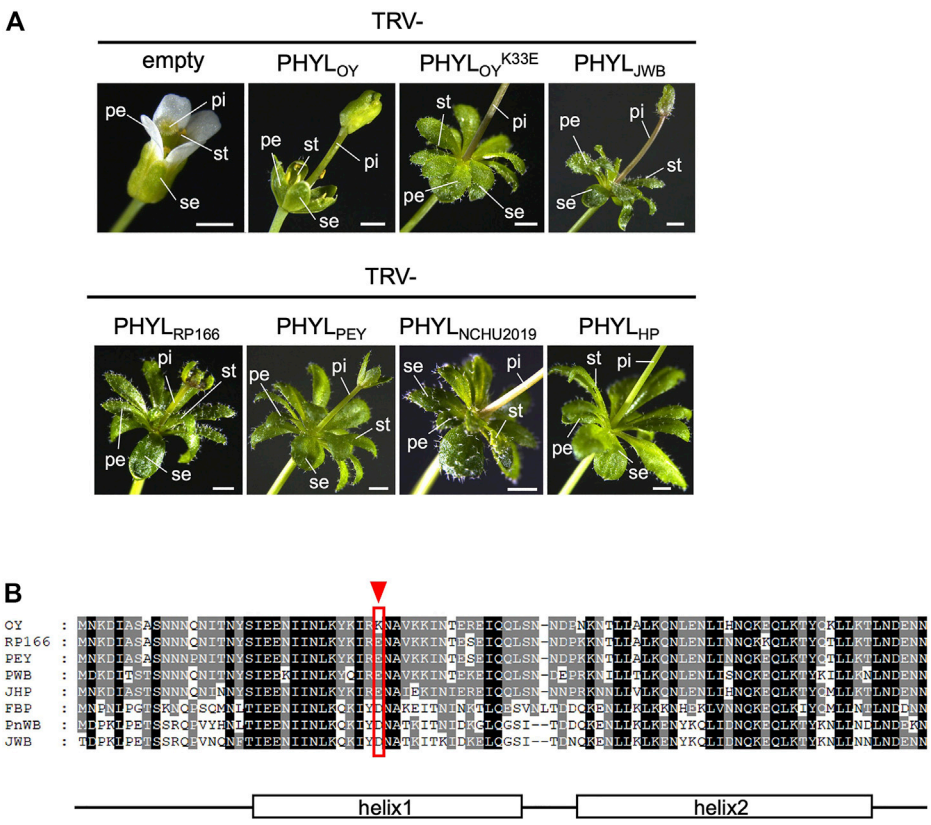


FIGURE 5
(A) Functional analysis of *phyllogens* associated with PMUs and PHYL$_{OY}^{K33E}$. *Arabidopsis thaliana* plants were infected with the tobacco rattle virus (TRV) vectors carrying *phyllogens* from either the phyl-A or phyl-D group. The following floral organs are shown: sepals (se), petals (pe), stamens (st), and pistils (pi). White scale bar: 1 mm. **(B)** Amino acid alignment of the secreted regions of PHYL$_{OY}$ and the phyl-A or phyl-D group phyllogens, which induce severe phyllody. The red arrowhead and border indicate a unique polymorphism at position 33 of PHYL$_{OY}$. Consensus secondary structure elements of phyllogens are shown (Iwabuchi et al., 2019).

low sequence identities (>32%), except for the closely related HP and RhY strains, and JWB-nky and Hebei-2018 strains (Supplementary Table S7; Figure 2B). Additionally, several PMU-associated genes, including *hflB* in the OY strain and *smc* in the HP, RhY, and PaWB-Japan strains, were truncated by premature stop codons due to frameshifts or single-nucleotide polymorphisms (Figure 1B). Moreover, the RP166, OY, PaWB-Japan, PaWB-China, RhY, and HP strains lacked PMU-associated genes downstream of *phyllogens* (e.g., *tmk*, *dnaB*, *dnaG*, and *tra5*; Figure 1B). These results indicate that although *phyllogens* are highly conserved in both types of PMU, *phyllogen* flanking genes have accumulated mutations, undergone pseudogenesis, and sometimes been lost altogether.

The phylogenetic relationships among '*Ca*. P. asteris' strains containing *phyllogens* were complex

Next, we focused on relationships among *phyllogens* in '*Ca*. P. asteris' because sequence information and genetic diversity were richer in '*Ca*. P. asteris' than in other phytoplasma species (Iwabuchi et al., 2020). Although many strains related to '*Ca*. P. asteris' had phyl-A group *phyllogens*, several strains had phyl-D group *phyllogens*.

First, intraspecies evolutionary relationships among strains with *phyllogens* were assessed by analyzing single-copy genes present in all whole/draft genome sequences. In total, 16 strains related to '*Ca*. P. asteris' were analyzed, with a '*Ca*. P. meliae' strain used as an outgroup. A WBD strain, which had recently been proposed for reclassification to '*Ca*. P. tritici' (Zhao et al., 2021), was also used in this analysis. The resulting phylogenetic tree showed that strains related to '*Ca*. P. asteris' formed two clades (Figure 3A). The minor clade comprised the AY-WB and WBD strains, each of which contained phyl-A group *phyllogens*. The major clade comprised four subclades and the TBZ1 strain. Two subclades comprised the strains with phyl-D group *phyllogens*; one of these subclades comprised the HP, RhY, and MD-China strains, and the other comprised the PaWB-Japan and PaWB-China strains. No *phyllogens* were found in the draft genome of the SW86 strain, which belonged to the former subclade. Another subclade comprised the strains containing phyl-A group *phyllogens* (i.e., the CYP, RP166, DY 2014, and OY strains). The final subclade comprised strains with complete (De Villa and M3 strains) or draft (LD1 strain) genomes that lacked *phyllogens*. The TBZ1 strain that contained a phyl-A group *phyllogen* did not belong to any of these subclades. These results indicate that the strains with phyl-D group *phyllogens* formed two subclades in '*Ca*. P. asteris', while the strains with phyl-A group *phyllogens*

formed one clade and one subclade. They also indicate that although strains related to '*Ca*. P. asteris' retained the same group of *phyllogens*, at least at the subclade level, the overall relationships were complex.

PMUs were inserted into different genomic regions in different '*Ca*. P. asteris' subgroups

The positions of the mobile elements within genomes may provide important evolutionary information. Therefore, we analyzed the genomic locations of PMUs in strains related to '*Ca*. P. asteris,' based on their flanking genomic regions. First, we analyzed the genomic locations of the type 1 PMUs. The PMUs and surrounding genomic regions in the OY, DY 2014, RP166, and AY-WB strains are shown in Figure 3B; Supplementary Figure S6B. The type 1 PMUs of the OY, DY 2014, and RP166 strains, which belong to the same subclade as '*Ca*. P. asteris' (Figure 3A), are flanked by several single-copy genes. These genes include *dppD*, *oppF*, *ibpA*, and *pacL* in the upstream region, and *nrdE*, *nrdF*, *mscL*, and *mdlB* in the downstream region (Figure 3B). The syntenies of these single-copy genes were also conserved in the other '*Ca*. P. asteris' genomes, but there were no *phyllogens* in the corresponding regions, except in the HP and RhY strains. However, type 2 PMUs harboring *phyllogens* were found in these corresponding genomic regions. The type 1 PMU in the AY-WB strain was in a different genomic region, and was flanked by several single-copy genes including *pdhC*, *acoL*, *tatD*, and *plsX* in the upstream region and *rpsD*, *mgtA*, *degV*, and *tsaD* in the downstream region (Supplementary Figure S6B). The syntenies of these genes were conserved in other '*Ca*. P. asteris' genomes, although no type 1 PMUs or *phyllogens* were found. These results indicate that the type 1 PMUs were inserted into different regions in the AY-WB and OY/DY2014/RP166 strains. Next, we analyzed the genomic locations of the type 2 PMUs. The PMUs and surrounding genomic regions in the PaWB-Japan and PaWB-China strains, which belong to the same subgroup, are shown in Figure 3C. These PMUs were delimited by a complete and truncated *tra5* gene downstream and upstream of the *phyllogens*, respectively (Bai et al., 2006). Several single-copy genes were also found near the PMUs, but these genes were different from those surrounding the type 1 PMUs; they included *rpsP*, *prfB*, *secA*, and *ypgC* in the upstream region and *tengu*, *glnQ*, and *artl* in the downstream region (Figure 3C). Although the syntenies of these single-copy genes were also conserved in other '*Ca*. P. asteris' genomes, there were no type 2 PMUs or *phyllogens*. The genomic region between the *ypgC* and *tengu* genes in the OY, DY 2014, De Villa, M3, and WBD strains was <13 kbp, which was much shorter than in the PaWB-Japan (35 kbp) and PaWB-China (67 kbp) strains. Several PMU-associated genes were found in this genomic region in the RP166 and AY-WB strains; however, the syntenies of these genes differed from those of the type 2 PMUs (Figure 3C). On the other hand, the type 2 PMUs in the HP and RhY strains were flanked by several single-copy genes different from those present in the PaWB strains (Figure 3B). The phyl-D group *phyllogen* present in the MD-China strain was also flanked by single-copy genes different from the genes flanking other *phyllogens* (Supplementary Figure S6). These results indicate that type 2 PMUs were inserted into different regions in the PaWB-Japan/PaWB-China, MD-China, and HP/RhY strains.

To correlate the inserted type 2 PMUs with other genomes at the nucleotide level, the upstream and downstream sequences of the PMUs in the PaWB strains were aligned with the corresponding genomic regions in the OY, De Villa, and DY2014 strains. The alignments identified regions of approximately 350 bp at both ends of the insertions that contained inverted repeat sequences, which is characteristic of transposon insertion (Szeverényi et al., 2003), and these inverted repeats were conserved between the PaWB strains (Figure 4; Supplementary Figure S7). Although these inverted repeat sequences were not identical in the other three strains, there were nearby sequences with strong similarity among these genomes (Figure 4), suggesting that PMU boundaries were located near the inverted repeats. These results indicate that a type 2 PMU was inserted at this position in the common ancestor of the PaWB strains.

Functions of PMU-associated phyllogens were conserved

Comparative analyses revealed conservation of amino acid sequences of the PMU-associated phyllogens. Thus, we examined functional conservation of the phyllogens by testing the capacity of various phyllogens to induce phyllody. We tested the phyl-A group phyllogens of RP166 ($PHYL_{RP166}$) and the PEY strains ($PHYL_{PEY}$), as well as the phyl-D group phyllogens of NCHU2019 ($PHYL_{NCHU2019}$) and the HP strains ($PHYL_{HP}$). We used the $PHYL_{OY}$ phyllogen of the OY strain (phyl-A group) and the $PHYL_{JWB}$ phyllogen of the JWB strain (phyl-D group) as controls known to induce phyllody (Iwabuchi et al., 2020). Each phyllogen was expressed in *A. thaliana* using a TRV-based gene expression vector system. All of the tested phyllogens converted sepals, petals, and stamens into leaf-like structures with stellate trichomes, changed pistils into secondary vegetative shoot-like structures, and enlarged flowers (Figure 5A), as did the $PHYL_{JWB}$ phyllogen control. These results showed that the functions of the type 1 and 2 PMU-associated phyllogens were as highly conserved as their sequences.

Interestingly, in the $PHYL_{OY}$ expression test, the stamens of most flowers were not converted into leaf-like structures, the other floral organs had fewer trichomes, and the flowers did not enlarge (Figure 5A). This suggested that $PHYL_{OY}$ exhibited a reduced capacity to induce phyllody. Sequence comparisons revealed that $PHYL_{OY}$ had a unique polymorphism at position 33 (lysine; Figure 5B). A $PHYL_{OY}$ mutant with a reciprocal substitution at this position (glutamate; $PHYL_{OY}^{K33E}$) induced phyllody to the same extent as the other homologs (Figure 5A). Therefore, this amino acid is important for strong induction of phyllody.

Discussion

Interspecies acquisition of *phyllogens via* horizontal transfer of PMUs

The apparent differences between the evolutionary histories of *phyllogens* and 16S rRNA genes in phytoplasmas suggest that *phyllogens* may have undergone HGT (Iwabuchi et al., 2020; Kalla et al., 2021). However, the mechanisms involved were previously unclear due to the lack of *phyllogen* flanking-region sequence data. In this study, we determined draft genomes for three strains and

analyzed *phyllogen* flanking regions from 17 phytoplasma strains related to six different species. We found that most *phyllogens* were associated with PMUs (Figure 1). Furthermore, *phyllogen* groupings correlated closely with the types of PMU (Figure 1). These results indicate that PMUs drive the horizontal transfer of distinct groups of *phyllogens* between phytoplasma species and strains. This study demonstrates that the symptom-determinant effectors of phytoplasmas are evolutionarily correlated with the types of PMU rather than the phytoplasma genomes.

In addition to phyllogens, three other effectors (TENGU, SAP11, and SAP05) have been linked experimentally with phytoplasma virulence (Hoshi et al., 2009; Sugawara et al., 2013; Minato et al., 2014; Cho et al., 2019; Huang et al., 2021). Among these, the SAP11 and SAP05 genes are also located near PMUs; they are conserved in various species (Sugio and Hogenhout., 2012; Cho et al., 2019; Huang et al., 2022) and have phylogenetic trees that differ from those of phytoplasmas (Chang et al., 2018; Huang et al., 2022). These observations suggest that PMU-mediated HGT may play a major role in the transfer of virulence genes across phytoplasma species.

Many studies have reported HGT of bacterial virulence genes based on genomic analyses of individual species (Ma et al., 2006; McCann and Guttman, 2008; Drew et al., 2021), but few have shown how HGT can occur *via* transposable elements (such as transposons) by comparing the genomes of different bacterial species. Thus, this study provides new insight into the mechanisms of HGT of virulence genes.

Phyllogen sequences and functions are conserved but PMU sequences may deteriorate

Genetic elements may become immobilized due to inactivation or deletion of genes necessary for their transfer (Dobrindt et al., 2004). Degeneration of PMUs has also been reported in several phytoplasma strains, including AY-WB (Bai et al., 2006; Ku et al., 2013). In this study, some PMUs harboring *phyllogens* lacked *tra5* (Figure 1B). In addition, some PMU genes had accumulated mutations, were truncated, or had been lost altogether (Figures 1B, 2), including genes with putative roles in DNA recombination, replication, and transposition (Arashida et al., 2008a; Ku et al., 2013). Therefore, most of these PMUs may begin to lose their capacity for transposition after the *phyllogens* have been acquired. However, the type 2 PMUs of the PaWB-Japan and PaWB-China strains retained their *tra5* genes and inverted repeat-like sequences at both ends (Figures 3, 4), suggesting that they had also retained their capacity for transposition.

Despite the deterioration of PMU-associated genes, *phyllogens* from different species had highly conserved sequences, especially *phyllogens* from the phyl-A and -D groups (Figure 2). Furthermore, phyllogen functions were conserved (Figure 5). Several *phyllogens* were not located within PMUs (Figure 1A; Chung et al., 2013; Luo et al., 2022) and their flanking regions exhibited considerable variation (data not shown), suggesting that after *phyllogen* transfer, PMU sequences may deteriorate completely. Therefore, *phyllogens* may become fixed in many different phytoplasma genomes after their acquisition, suggesting that they are crucial for their hosts. Previous studies have shown that phyllody symptoms facilitate phytoplasma accumulation/localization within host plants (Arashida et al., 2008b; Su et al., 2011), and phyllogens can attract insect vectors (MacLean et al., 2014). Although further studies are needed to investigate the roles of phyllody symptoms in host adaptability, the fact that *phyllogens* have been strongly conserved throughout evolution suggests that they are critical for phytoplasma survival.

Evolutionary history of *phyllogen* acquisition in '*Ca*. P. asteris'

To date, three different groups of *phyllogens* have been identified in strains related to '*Ca*. P. asteris' (phyl-A, -B, and -D; Iwabuchi et al., 2020). In this study, we found that the acquisition of these *phyllogens* was a complex process, at least for phyl-A and phyl-D. In '*Ca*. P. asteris,' we identified one subclade comprising strains that retained phyl-A *phyllogens* and two subclades comprising strains that retained phyl-D *phyllogens* (Figure 3A). In the former subclade, the PMUs retaining phyl-A were found in the same region of the genomes (Figure 3), suggesting that these PMUs were acquired by their common ancestor. The same region also surrounded PMUs of the HP and RhY strains (Figure 3). Therefore, this region may be a hot spot for genome rearrangement in phytoplasmas, as described by a previous study (Arashida et al., 2008b). PMUs retaining phyl-D were found in different genomic regions in the PaWB-Japan/PaWB-China, MD-China, and HP/RhY strains (Figure 3; Supplementary Figure S6). This observation suggests that PMUs harboring *phyllogens* either moved frequently within the genome or were acquired separately. The PaWB strains had complete PMUs retaining phyl-D, suggesting that these were acquired more recently than the PMUs in the HP and RhY strains. Therefore, the latter hypothesis seems more plausible. We could not determine the acquisition order of the phyl-A and phyl-D group *phyllogens* in '*Ca*. P. asteris' due to insufficient genomic data. Therefore, further studies of *phyllogen* flanking regions and their genomic positions are needed.

HGT may occur between organisms that share an ecological niche, irrespective of their phylogenetic relationships (Polz et al., 2013). Phylogenetic analyses have shown that all of the phyl-D group *phyllogens* of '*Ca*. P. asteris' are most closely related to *phyllogens* of '*Ca*. P. ziziphi' (Iwabuchi et al., 2020; Figure 1A). Furthermore, several lines of evidence indicate that strains related to '*Ca*. P. asteris' and '*Ca*. P. ziziphi' share common ecological niches. First, the distribution of '*Ca*. P. ziziphi' is limited to eastern and southern Asia (Jung et al., 2003; Rao et al., 2017), as is the distribution of strains related to '*Ca*. P. asteris' that retain phyl-D (Kakizawa et al., 2006; Takinami et al., 2013; Luo et al., 2022). Second, co-infection of '*Ca*. P. asteris' and '*Ca*. P. ziziphi' has been reported in jujube plants (Sun et al., 2013). Third, *Hishimonus sellatus*, a vector of '*Ca*. P. ziziphi,' can transmit two strains related to '*Ca*. P. asteris' that retain phyl-D (RhY and MD-China; Tanaka et al., 2000; Kusunoki et al., 2002). Therefore, type 2 PMUs in '*Ca*. P. asteris' may be acquired from '*Ca*. P. ziziphi' in the same ecological niche. Further accumulation of information on the synteny of PMUs will reveal how PMUs and

symptom-determinant genes located in them are shared among phytoplasmas in the same niche.

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 below: https://www.ddbj.nig.ac.jp/, BSDA00000000; https://www.ddbj.nig.ac.jp/, BSCX00000000; https://www.ddbj.nig.ac.jp/, BSCY00000000; https://www.ddbj.nig.ac.jp/, LC740440; https://www.ddbj.nig.ac.jp/, LC740441; https://www.ddbj.nig.ac.jp/, LC740442; https://www.ddbj.nig.ac.jp/, LC740443; https://www.ddbj.nig.ac.jp/, LC740444; https://www.ddbj.nig.ac.jp/, LC740445; https://www.ddbj.nig.ac.jp/, LC740446; https://www.ddbj.nig.ac.jp/, LC740447; https://www.ddbj.nig.ac.jp/, LC740448; https://www.ddbj.nig.ac.jp/, LC740449; https://www.ddbj.nig.ac.jp/, LC740450; https://www.ddbj.nig.ac.jp/, LC740451; https://www.ddbj.nig.ac.jp/, LC740452; https://www.ddbj.nig.ac.jp/, LC740453.

Author contributions

RT, NI, and KM designed the research. YK, TN, KO, SN, and YY supervised the experiments. RT performed most experiments. MS provided technical assistance in the experiments. RT wrote the initial draft of the article. MS complemented the writing. NI, YK, TN, KM, KO, SN, and YY critically revised and edited the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) through Grants-in-Aid for Scientific Research (nos. 19K15840, 20H02991, 20J23168, 20K22562, 20KK0132, 21H04722, 21K14847, 21K14853, and 21K19239).

Acknowledgments

We are grateful to Drs T. Shiomi (Department of Biological Safety, National Institute for Agro-Environmental Sciences, Ibaraki, Japan), T. Usugi (Plant Protection Division, National Agricultural Research Center, Ibaraki, Japan), N. Nishimura and T. Tsuchizaki (Department of Agribusiness, Koibuchi College of Agriculture and Nutrition, Ibaraki, Japan), and A. Bertaccini (DipSA, University of Bologna, Bologna, Italy) for providing phytoplasma DNA 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/fgene.2023.1132432/full#supplementary-material

References

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

Arashida, R., Kakizawa, S., Hoshi, A., Ishii, Y., Jung, H. Y., Kagiwada, S., et al. (2008a). Heterogeneic dynamics of the structures of multiple gene clusters in two pathogenetically different lines originating from the same phytoplasma. *DNA Cell Biol.* 27, 209–217. doi:10.1089/dna.2007.0654

Arashida, R., Kakizawa, S., Ishii, Y., Hoshi, A., Jung, H. Y., Kagiwada, S., et al. (2008b). Cloning and characterization of the antigenic membrane protein (AMP) gene and *in situ* detection of AMP from malformed flowers infected with Japanese Hydrangea phyllody phytoplasma. *Phytopathology* 98, 769–775. doi:10.1094/PHYTO-98-7-0769

Bai, X., Zhang, J., Ewig, A., Miller, S. A., Radek, A. J., Shevchenko, D. V., et al. (2006). Living with genome instability: The adaptation of phytoplasmas to diverse environments of their insect and plant hosts. *J. Bacteriol.* 188, 3682–3696. doi:10.1128/JB.188.10.3682-3696.2006

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinform* 10, 421. doi:10.1186/1471-2105-10-421

Chang, S. H., Tan, C. M., Wu, C. T., Lin, T. H., Jiang, S. Y., Liu, R. C., et al. (2018). Alterations of plant architecture and phase transition by the phytoplasma virulence factor SAP11. *J. Exp. Bot.* 69, 5389–5401. doi:10.1093/jxb/ery318

Cho, S. T., Kung, H. J., Huang, W., Hogenhout, S. A., and Kuo, C. H. (2020). Species boundaries and molecular markers for the classification of 16SrI phytoplasmas inferred by genome analysis. *Front. Microbiol.* 11, 1531. doi:10.3389/fmicb.2020.01531

Cho, S. T., Lin, P., and Kuo, C. H. (2019). Genomic characterization of the periwinkle leaf yellowing (PLY) phytoplasmas in Taiwan. *Front. Microbiol.* 10, 2194. doi:10.3389/fmicb.2019.02194

Chung, W. C., Chen, L. L., Lo, W. S., Lin, C. P., and Kuo, C. H. (2013). Comparative analysis of the peanut witches'-broom phytoplasma genome reveals horizontal transfer of potential mobile units and effectors. *PLoS One* 8, e62770. doi:10.1371/journal.pone.0062770

Cosentino, S., and Iwasaki, W. (2019). SonicParanoid: Fast, accurate and easy orthology inference. *Bioinformatics* 35, 149–151. doi:10.1093/bioinformatics/bty631

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* 2, 414–424. doi:10.1038/nrmicro884

Drew, G. C., Stevens, E. J., and King, K. C. (2021). Microbial evolution and transitions along the parasite–mutualist continuum. *Nat. Rev. Microbiol.* 19, 623–638. doi:10.1038/s41579-021-00550-7

Gilchrist, C. L. M., and Chooi, Y. H. (2021). Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 37, 2473–2475. doi:10.1093/bioinformatics/btab007

Gluck-Thaler, E., Cerutti, A., Perez-Quintero, A. L., Butchacas, J., Roman-Reyna, V., Madhavan, V. N., et al. (2020). Repeated gain and loss of a single gene modulates the evolution of vascular plant pathogen lifestyles. *Sci. Adv.* 6, eabc4516. doi:10.1126/sciadv.abc4516

Hooykaas, P. J. J., Snijdewint, F. G. M., and Schilperoort, R. A. (1982). Identification of the sym plasmid of *Rhizobium leguminosarum* strain 1001 and its transfer to and expression in other Rhizobia and *Agrobacterium tumefaciens*. *Plasmid* 8, 73–82. doi:10.1016/0147-619x(82)90042-7

Hoshi, A., Oshima, K., Kakizawa, S., Ishii, Y., Ozeki, J., Hashimoto, M., et al. (2009). A unique virulence factor for proliferation and dwarfism in plants identified from a phytopathogenic bacterium. *Proc. Natl. Acad. Sci. U. S. A.* 106, 6416–6421. doi:10.1073/pnas.0813038106

Huang, C. T., Cho, S. T., Lin, Y. C., Tan, C. M., Chiu, Y. C., Yang, J. Y., et al. (2022). Comparative genome analysis of '*Candidatus* Phytoplasma luffae' reveals the influential roles of potential mobile units in phytoplasma evolution. *Front. Microbiol.* 13, 773608. doi:10.3389/fmicb.2022.773608

Huang, W., MacLean, A. M., Sugio, A., Maqbool, A., Busscher, M., Cho, S. T., et al. (2021). Parasitic modulation of host development by ubiquitin-independent protein degradation. *Cell* 184, 5201–5214.e12. doi:10.1016/j.cell.2021.08.029

Iwabuchi, N., Kitazawa, Y., Maejima, K., Koinuma, H., Miyazaki, A., Matsumoto, O., et al. (2020). Functional variation in phyllogen, a phyllody-inducing phytoplasma effector family, attributable to a single amino acid polymorphism. *Mol. Plant. Pathol.* 21, 1322–1336. doi:10.1111/mpp.12981

Iwabuchi, N., Maejima, K., Kitazawa, Y., Miyatake, H., Nishikawa, M., Tokuda, R., et al. (2019). Crystal structure of phyllogen, a phyllody-inducing effector protein of phytoplasma. *Biochem. Biophys. .Res. Commun.* 513, 952–957. doi:10.1016/j.bbrc.2019.04.060

Jomantiene, R., Zhao, Y., and Davis, R. E. (2007). Sequence-variable mosaics: Composites of recurrent transposition characterizing the genomes of phylogenetically diverse phytoplasmas. *DNA Cell. Biol.* 26, 557–564. doi:10.1089/dna.2007.0610

Jung, H. Y., Sawayanagi, T., Kakizawa, S., Nishigawa, H., Wei, W., Oshima, K., et al. (2003). '*Candidatus* Phytoplasma ziziphi', a novel phytoplasma taxon associated with jujube witches'-broom disease. *Int. J. Syst. Evol. Microbiol.* 53, 1037–1041. doi:10.1099/ijs.0.02393-0

Kakizawa, S., Oshima, K., Jung, H. Y., Suzuki, S., Nishigawa, H., Arashida, R., et al. (2006). Positive selection acting on a surface membrane protein of the plant-pathogenic phytoplasmas. *J. Bacteriol.* 188, 3424–3428. doi:10.1128/JB.188.9.3424-3428.2006

Kalla, S., Verma, P., and Lakhanpaul, S. (2021). Understanding the evolution of SAP54, a phyllody associated phytoplasma effector gene through phylogenetic network analysis. *Phytopathog. Mollicutes.* 11, 36–44. doi:10.5958/2249-4677.2021.00005.0

Keeling, P. J., and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9, 605–618. doi:10.1038/nrg2386

Kitazawa, Y., Iwabuchi, N., Himeno, M., Sasano, M., Koinuma, H., Nijo, T., et al. (2017). Phytoplasma-conserved phyllogen proteins induce phyllody across the Plantae by degrading floral MADS domain proteins. *J. Exp. Bot.* 68, 2799–2811. doi:10.1093/jxb/erx158

Kitazawa, Y., Iwabuchi, N., Maejima, K., Sasano, M., Matsumoto, O., Koinuma, H., et al. (2022). A phytoplasma effector acts as a ubiquitin-like mediator between floral MADS-box proteins and proteasome shuttle proteins. *Plant Cell* 34, 1709–1723. doi:10.1093/plcell/koac062

Ku, C., Lo, W. S., and Kuo, C. H. (2013). Horizontal transfer of potential mobile units in phytoplasmas. *Mob. Genet. Elem.* 3, e26145. doi:10.4161/mge.26145

Kube, M., Schneider, B., Kuhl, H., Dandekar, T., Heitmann, K., Migdoll, A. M., et al. (2008). The linear chromosome of the plant-pathogenic mycoplasma '*Candidatus* Phytoplasma Mali. *BMC Genom* 9, 306. doi:10.1186/1471-2164-9-306

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

Kusunoki, M., Shiomi, T., Kobayashi, M., Okudaira, T., Ohashi, A., and Nohira, T. (2002). A leafhopper (*Hishimonus sellatus*) transmits phylogenetically distant phytoplasmas: Rhus yellows and hovenia witches' broom phytoplasma. *J. Gen. Plant. Pathol.* 68, 147–154. doi:10.1007/PL00013068

Luo, L., Zhang, X., Meng, F., Wang, Y., Zhou, Y., and Liu, J. (2022). Draft genome sequences resources of mulberry dwarf phytoplasma strain MDGZ-01 associated with mulberry yellow dwarf (MYD) diseases. *Plant Dis.* 106, 2239–2242. doi:10.1094/PDIS-12-21-2671-A

Ma, W., Dong, F. F. T., Stavrinides, J., and Guttman, D. S. (2006). Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS. Genet.* 2, e209–e2142. doi:10.1371/journal.pgen.0020209

MacLean, A. M., Orlovskis, Z., Kowitwanich, K., Zdziarska, A. M., Angenent, G. C., Immink, R. G. H., et al. (2014). Phytoplasma effector SAP54 hijacks plant reproduction by degrading MADS-box proteins and promotes insect colonization in a RAD23-dependent manner. *PLoS Biol.* 12, e1001835. doi:10.1371/journal.pbio.1001835

MacLean, A. M., Sugio, A., Makarova, O. V., Findlay, K. C., Grieve, V. M., Toth, R., et al. (2011). Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in Arabidopsis plants. *Plant Physiol.* 157, 831–841. doi:10.1104/pp.111.181586

Maejima, K., Iwai, R., Himeno, M., Komatsu, K., Kitazawa, Y., Fujita, N., et al. (2014). Recognition of floral homeotic MADS-domain transcription factors by a phytoplasmal effector, phyllogen, induces phyllody. *Plant J.* 78, 541–554. doi:10.1111/tpj.12495

Marcone, C. (2014). Molecular biology and pathogenicity of phytoplasmas. *Ann. Appl. Biol.* 165, 199–221. doi:10.1111/aab.12151

McCann, H. C., and Guttman, D. S. (2008). Evolution of the type III secretion system and its effectors in plant–microbe interactions. *New. Phytol.* 177, 33–47. doi:10.1111/j.1469-8137.2007.02293.x

Melnyk, R. A., Hossain, S. S., and Haney, C. H. (2019). Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated Pseudomonas. *ISME J.* 13, 1575–1588. doi:10.1038/s41396-019-0372-5

Minato, N., Himeno, M., Hoshi, A., Maejima, K., Komatsu, K., Takebayashi, Y., et al. (2014). The phytoplasmal virulence factor TENGU causes plant sterility by downregulating of the jasmonic acid and auxin pathways. *Sci. Rep.* 4, 7399. doi:10.1038/srep07399

Muhire, B. M., Varsani, A., and Martin, D. P. (2014). Sdt: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE* 9, e108277. doi:10.1371/journal.pone.0108277

Music, M. S., Samarzija, I., Hogenhout, S. A., Haryono, M., Cho, S. T., and Kuo, C. H. (2019). The genome of '*Candidatus* Phytoplasma solani' strain SA-1 is highly dynamic and prone to adopting foreign sequences. *Syst. Appl. Microbiol.* 42, 117–127. doi:10.1016/j.syapm.2018.10.008

Namba, S., Oyaizu, H., Kato, S., Iwanami, S., and Tsuchizaki, T. (1993). Phylogenetic diversity of phytopathogenic mycoplasmalike organisms. *Int. J. Syst. Bacteriol.* 43, 461–467. doi:10.1099/00207713-43-3-461

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi:10.1093/dnares/dsn027

Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H. Y., Wei, W., Suzuki, S., et al. (2004). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat. Genet.* 36, 27–29. doi:10.1038/ng1277

Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 170–175. doi:10.1016/j.tig.2012.12.006

Rao, G. P., Priya, M., Thorat, V., Manimekalai, R., Tiwari, K., and Yadav, A. (2017). A century progress of research on phytoplasma diseases in India. *Phytopathog. Mollicutes.* 7, 1–38. doi:10.5958/2249-4677.2017.00001.9

Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Comput. Biol.* 11, e1004095. doi:10.1371/journal.pcbi.1004095

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Soucy, S., Huang, J., and Gogarten, J. (2015). Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* 16, 472–482. doi:10.1038/nrg3962

Su, Y. T., Chen, J. C., and Lin, C. P. (2011). Phytoplasma-induced floral abnormalities in Catharanthus roseus are associated with phytoplasma accumulation and transcript repression of floral organ identity genes. *Mol. Plant Microbe Interact.* 24, 1502–1512. doi:10.1094/MPMI-06-11-0176

Sugawara, K., Honma, Y., Komatsu, K., Himeno, M., Oshima, K., and Namba, S. (2013). The alteration of plant morphology by small peptides released from the proteolytic processing of the bacterial peptide TENGU. *Plant Physiol.* 162, 2005–2014. doi:10.1104/pp.113.218586

Sugio, A., and Hogenhout, S. A. (2012). The genome biology of phytoplasma: Modulators of plants and insects. *Curr. Opin. Microbiol.* 15, 247–254. doi:10.1016/j.mib.2012.04.002

Sugio, A., Kingdom, H. N., MacLean, A. M., Grieve, V. M., and Hogenhout, S. A. (2011). Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 108, E1254–E1263. doi:10.1073/pnas.1105664108

Sun, X. C., Mou, H. Q., Li, T. T., Tian, Q., and Zhao, W. J. (2013). Mixed infection of two groups (16SrI & V) of phytoplasmas in a single jujube tree in China. *J. Phytopathol.* 161, 661–665. doi:10.1111/jph.12101

Szeverényi, I., Nagy, Z., Farkas, T., Olasz, F., and Kiss, J. (2003). Detection and analysis of transpositionally active head-to-tail dimers in three additional *Escherichia coli* IS elements. *Microbiology* 149, 1297–1310. doi:10.1099/mic.0.26121-0

Takahashi, S., Komatsu, K., Kagiwada, S., Ozeki, J., Mori, T., Hirata, H., et al. (2006). The efficiency of interference of potato virus X infection depends on the target gene. *Virus Res.* 116, 214–217. doi:10.1016/j.virusres.2005.11.002

Takinami, Y., Maejima, K., Takahashi, A., Keima, T., Shiraishi, T., Okano, Y., et al. (2013). First report of '*Candidatus* Phytoplasma asteris' infecting hydrangea showing phyllody in Japan. *J. Gen. Plant Pathol.* 79, 209–213. doi:10.1007/s10327-013-0445-7

Tanaka, M., Osada, S., Matsuda, I., and Marks, R. (2000). Transmission of rhus (*Rhus javanica* L.) yellows by *Hishimonus sellatus* and host range of the causal phytoplasma. *J. Gen. Plant Pathol.* 66, 323–326. doi:10.1046/j.1365-2230.2000.00653.x

Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018). Dfast: A flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* 34, 1037–1039. doi:10.1093/bioinformatics/btx713

Toruño, T. Y., Musić, M. S., Simi, S., Nicolaisen, M., and Hogenhout, S. A. (2010). Phytoplasma PMU1 exists as linear chromosomal and circular extrachromosomal elements and has enhanced expression in insect vectors compared with plant hosts. *Mol. Microbiol.* 77, 1406–1415. doi:10.1111/j.1365-2958.2010.07296.x

Tran-Nguyen, L. T., Kube, M., Schneider, B., Reinhardt, R., and Gibb, K. S. (2008). Comparative genome analysis of '*Candidatus* Phytoplasma australiense' (subgroup *tuf*-Australia I; *rp*-A) and '*Ca*. Phytoplasma asteris' strains OY-M and AY-WB. *J. Bacteriol.* 190, 3979–3991. doi:10.1128/JB.01301-07

Wang, J., Song, L., Jiao, Q., Yang, S., Gao, R., Lu, X., et al. (2018). Comparative genome analysis of jujube witches'-broom Phytoplasma, an obligate pathogen that causes jujube witches'-broom disease. *BMC Genom* 19, 689. doi:10.1186/s12864-018-5075-1

Wei, W., Davis, R. E., Jomantiene, R., and Zhao, Y. (2008). Ancient, recurrent phage attacks and recombination shaped dynamic sequence-variable mosaics at the root of phytoplasma genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 11827–11832. doi:10.1073/pnas.0805237105

Yang, C. Y., Huang, Y. H., Lin, C. P., Lin, Y. Y., Hsu, H. C., Wang, C. N., et al. (2015). MicroRNA396-targeted *SHORT VEGETATIVE PHASE* is required to repress flowering and is related to the development of abnormal flower symptoms by the phyllody symptoms1 effector. *Plant Physiol.* 168, 1702–1716. doi:10.1104/pp.15.00307

Zhao, Y., Wei, W., Davis, R. E., Lee, I. M., and Bottner-Parker, K. D. (2021). The agent associated with blue dwarf disease in wheat represents a new phytoplasma taxon, '*Candidatus* Phytoplasma tritici. *Int. J. Syst. Evol.* 71, 004604. doi:10.1099/ijsem.0.004604

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 15 May 2023
DOI 10.3389/fmicb.2023.1179966

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Xiaogang Xu,
Fudan University, China
Ahmed M. Moustafa,
University of Pennsylvania, United States

*CORRESPONDENCE
Xin Lu
luxin@icdc.cn
Zhenpeng Li
lizhenpeng@icdc.cn
Biao Kan
kanbiao@icdc.cn

RECEIVED 05 March 2023
ACCEPTED 27 April 2023
PUBLISHED 15 May 2023

CITATION
Zhao W, Zeng W, Pang B, Luo M, Peng Y, Xu J, Kan B, Li Z and Lu X (2023) Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing. *Front. Microbiol.* 14:1179966. doi: 10.3389/fmicb.2023.1179966

COPYRIGHT
© 2023 Zhao, Zeng, Pang, Luo, Peng, Xu, Kan, Li and Lu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing

Wenxuan Zhao[1,2], Wei Zeng[1,3], Bo Pang[1], Ming Luo[4], Yao Peng[1], Jialiang Xu[5], Biao Kan[1,3]*, Zhenpeng Li[1]*and Xin Lu[1]*

[1]National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China, [2]Department of Health Statistics, School of Public Health, Shanxi Medical University, Jinzhong, Shanxi, China, [3]School of Public Health, Shandong University, Jinan, China, [4]Yulin Center for Disease Control and Prevention, Yulin, Shanxi, China, [5]School of Food and Chemical Engineering, Beijing Technology and Business University, Beijing, China

Introduction: Genome-based analysis is crucial in monitoring antibiotic-resistant bacteria (ARB)and antibiotic-resistance genes (ARGs). Short-read sequencing is typically used to obtain incomplete draft genomes, while long-read sequencing can obtain genomes of multidrug resistance (MDR) plasmids and track the transmission of plasmid-borne antimicrobial resistance genes in bacteria. However, long-read sequencing suffers from low-accuracy base calling, and short-read sequencing is often required to improve genome accuracy. This increases costs and turnaround time.

Methods: In this study, a novel ONT sequencing method is described, which uses the latest ONT chemistry with improved accuracy to assemble genomes of MDR strains and plasmids from long-read sequencing data only. Three strains of Salmonella carrying MDR plasmids were sequenced using the ONT SQK-LSK114 kit with flow cell R10.4.1, and de novo genome assembly was performed with average read accuracy (Q>10) of 98.9%.

Results and Discussion: For a 5-Mb-long bacterial genome, finished genome sequences with accuracy of >99.99% could be obtained at 75× sequencing coverage depth using Flye and Medaka software. Thus, this new ONT method greatly improves base-calling accuracy, allowing for the de novo assembly of high-quality finished bacterial or plasmid genomes without the need for short-read sequencing. This saves both money and time and supports the application of ONT data in critical genome-based epidemiological analyses. The novel ONT approach described in this study can take the place of traditional combination genome assembly based on short- and long-read sequencing, enabling pangenomic analyses based on high-quality complete bacterial and plasmid genomes to monitor the spread of antibiotic-resistant bacteria and antibiotic resistance genes.

KEYWORDS

ONT sequencing, accuracy, complete genome, bacteria, plasmid

1. Introduction

Microbial resistance to drugs has become a global issue of widespread concern (Nathan, 2020). The widespread emergence of antibiotic resistance, especially multidrug resistance (MDR), among bacterial strains that cause infections presents difficulties in clinical treatment. Acquired resistance spreads very rapidly compared with chromosomal mutations (Paterson and van Duin, 2017; Bengtsson-Palme et al., 2018). Plasmids are the most common vectors for horizontal gene transfer (San Millan, 2018; Vit et al., 2020). In the concept of One Health, it is crucial to investigate MDR plasmids, because their domain organization is critical to the spread of antimicrobial resistance genes (ARGs) among bacteria (Aslam et al., 2021). Therefore, to track the transmission of ARGs, accurate information on MDR plasmid genomes is essential (Bennett, 2008; Malhotra-Kumar et al., 2016; Jordt et al., 2020).

MDR mostly originates from the accumulation of resistance genes on plasmids (Nikaido, 2009), though resistance genes can also be carried on the chromosome. Identifying these genes and their accurate genomic localization using short-read sequencing data can be difficult (Partridge et al., 2009). To associate independent data with ARGs transmission events, pangenome clustering based on complete plasmid genomes can be applied to surveillance (Li et al., 2022). Short-read sequencing [such as from Illumina and MGI next-generation sequencing (NGS) technologies] has high base-calling accuracy, but the nature of the short reads means that only fragmented draft genomes can be obtained from such data. Instead, scientists would prefer to receive correct, complete genomes as their study advances (Cohen et al., 2019; Kathirvel et al., 2021). Shortly after the introduction of NGS, third-generation sequencing technologies (TGS) emerged, presented by two platforms, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), giving long and ultra-long sequencing reads, respectively, and enabling coverage of highly repetitive regions and structural variants. PacBio developed the first established single-molecule real-time sequencing technology in 2011 (Athanasopoulou et al., 2021). In 2014, high-throughput, long-read sequencing was made possible by ONT on a portable device MinION (Lu et al., 2016). However, compared with NGS, low base-call accuracy has limited the reliability of ONT data for critical genomic epidemiology tasks (Petersen et al., 2019; Foster-Nyarko et al., 2023). The widely accepted remedy was to use short-read NGS data for error correction of long-read sequence data (Senol Cali et al., 2019; Smith et al., 2020).

In this study, our main objectives were to produce a high-quality finished genome only by *de novo* assembly based on long-read sequencing and to offer solid evidence for resistance gene analysis, supporting its use in the genome-based epidemiological analyses. This approach offered a more efficient and cost-effective replacement for traditional methods that required both long- and short-read sequencing. First, the reference genome of the strains was generated. Then, the sequencing data were obtained using the new ONT SQK-LSK114 kit with flow cell R10.4.1, and use of long-read sequencing data only for *de novo* genome assembly of these strains, combined with the error correction of the data itself. Finally, the accuracy of these genomes was verified against the reference data. To evaluate the accuracy of our method, we assessed the single nucleotide variations (SNVs), insertions (INSs), and deletions (DELs), which are common *de novo* assembly errors (Boostrom et al., 2022). To compare the new sequencing method with the earlier sequencing methods, all the samples were sequenced using the SQK-LSK110 kit and the R9.4 flow cell. In comparison with the previous version (Sereika et al., 2022), sequencing quality was substantially improved in the latest ONT chemistry and has potential implications for monitoring the spread of antibiotic-resistant bacteria and antibiotic-resistant genes.

2. Materials and methods

2.1. Samples

Salmonella strains were collected from the surveillance of healthy people, and we constructed a strain bank from these strains. Antimicrobial susceptibility testing was conducted and interpreted using the broth microdilution method recommended by the Clinical and Laboratory Standards Institute. Three MDR *Salmonella* strains were recovered from the strain bank and isolated with *Salmonella* agar (CHROMagar Company, Paris, France); strains were identified using the Vitek-II system (bioMérieux, Lyon, France). The genomic DNA from each strains was extracted by boiling and freeze-thawing processes, and the resulting supernatant was recovered for use as the PCR template (Doyle et al., 2012; Ding et al., 2020; Fan et al., 2022).

2.2. Sequencing

Genomic DNA was extracted using the TIANamp Bacteria DNA Kit (TIANGEN, China) and quantified using a Qubit V4 Fluorometer (Thermo Fisher, United States). Sequencing libraries were prepared using Ligation Sequencing Kit V14 (SQK-LSK114, ONT) and sequenced using R10.4.1 flow cells (FLO-MIN114, ONT) on a GridION device (ONT) with MinKNOW v22.08.9 and super-accuracy base-calling mode selected. Other parameters were kept at their defaults. To compare the new sequencing method used in our study with the earlier sequencing methods, we sequenced all the samples using the SQK-LSK110 kit and the R9.4 flow cell.

NGS libraries were constructed using the MGIEasy FS DNA Library Prep Set (MGI, China) and sequenced on the MGISEQ-200RS sequencing platform (MGI).

2.3. Data analysis

Guppy v6.2.11 (Wick et al., 2019) was used to extract the bases from the downloaded fast5 data and turn them into standard fastq files. NanoPlot v1.20.0 was used to assess the level of sequencing quality (De Coster et al., 2018). NanoFilt v2.8 (De Coster et al., 2018) was used to remove sequences that were <1,000 bp long with quality value < 10. In addition, 50 bp were removed from the front and back ends of each clean data record.

It has been widely used in earlier studies and proved to be the most accurate method to assemble reference genomes utilizing short- and long-read sequencing (Petersen et al., 2019; Boostrom et al., 2022; Sereika et al., 2022). In our study, we utilized this

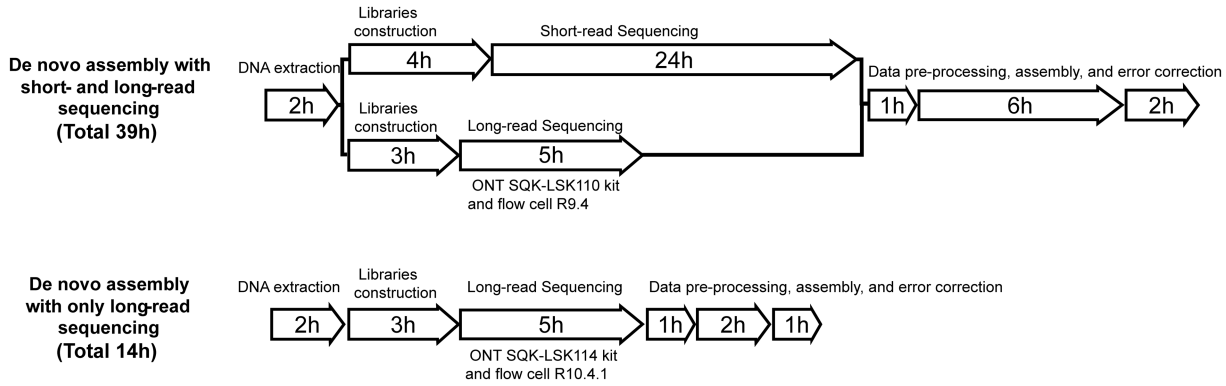


FIGURE 1
Time required to perform each method.

method to generate the reference genomes. Long-read data with 500 × depth and short read data with 500 × depth were used, and were assembled using the Unicycler hybrid assembler v0.4.8 (Wick et al., 2017). Pilon v1.24 (Walker et al., 2014) was used to polish these genomes. The obtained genomes were used as references for the following analyses.

Raw data packets were generated periodically during nanopore sequencing, typically at 6-min intervals (approximately 4,000 reads). The depth of coverage was based on packet size. After the clean data were generated, it was divided into depths: 1× (6 min), 5× (12 min), 10× (25 min), 20× (40 min), 30× (75 min), 50× (105 min), 75× (175 min), 100× (265 min), 150× (360 min), 200× (460 min), 250× (560 min), 300× (660 min), 350× (760 min), 400× (850 min), 450× (960 min), and 500× (1,050 min). The final depth of coverage was estimated based on the actual data size. The goal was to determine the saturation sequencing time point and depth.

Flye has proven to be the most effective tool for *de novo* genetic assembly (Boostrom et al., 2022). Therefore, the default parameters of Flye v2.8.2 (Kolmogorov et al., 2019) were used for *de novo* assembly, and QUAST v5.2.0 (Gurevich et al., 2013) was used to evaluate the quality of genome assembly. Finally, errors were corrected by applying Medaka v1.2.2 three times.[1] BLAST v2.11.0 was utilized to determine the identity of the fastq file in comparison to the reference genome. Snippy v4.4.5 was used to compare the assembled fasta file to the reference genome and obtain the number of SNPs, insertions, and deletions.[2] ARG genes were aligned using the ResFinder database. FastANI v1.33 (Jain et al., 2018) was used to calculate genome-wide average nucleotide identity between genomes. R v4.1.0 and BRIG v0.95 (Alikhan et al., 2011) were used to visualize the outcomes (Figure 1). For detailed usage instructions, please refer to the Supplementary material.

The complete sequences of all three strains have been deposited in the GenBank database, Their BioProject ID is PRJNA937772.

1 https://github.com/nanoporetech/medaka

2 https://github.com/tseemann/snippy

3. Results

3.1. Characteristics of sequences

In total, we used three MDR *Salmonella* strains (SA15303, SA14318, and SA17155) that have been shown by PCR to carry the *mcr*, *ndm*, and *tet* genes, respectively. After hybrid genome assembly from long and short reads, we used short reads for multiple rounds of polishing. The aim was to obtain a high-quality reference genome for each strain. The complete genomes for strains SA15303, SA14318, and SA17155 were 5.1, 4.7, and 5.2 Mb long, respectively. The plasmids of these strains were 29, 199, and 287 kb in size, and contained 4, 9, and 12 ARGs, respectively. Thus, the plasmid size as well as the resistance genes varied widely between the strains. There were more homopolymer regions on the larger plasmids (Table 1).

In experiments to test genome generation using only long-read sequencing data, we used a combination of the latest ONT kit v14 (SQK-LSK114) and flow cell R10.4.1 to obtain the raw data. After sequencing for 18 h, each flow cell generated about 500× data, with 39%–42% of the reads having lengths > 5 kb. Multiple reads of >60,000 bp were obtained (Figure 2A). Reads with quality value Q10 exceeded 85%, while those with Q20 exceeded 40% (Table 2).

We used an identical procedure to sequence the genomes of the three strains. The read lengths and Q-score distribution of the three sets of sequence data were similar (Figure 2A). The average length was 5 kb, and the average Q value was 16.70. We considered a Q value < 10 to represent low-quality reads and removed these (approximately 14% of the whole sequencing data) before assembly. When compared with the reference genome sequences, the average read accuracy was 98.9% (Figure 2B), with the lower quartile > 95%.

To compare the new sequencing method with the earlier sequencing methods, all the samples were sequenced using the SQK-LSK110 kit and the R9.4 flow cell. It was important to note that the new sequencing technique shown considerable increases in both quality and accuracy. More specifically, accuracy increased from 92% to 98% and Q20 increased from 13% to 42% (Table 2).

TABLE 1 Details of the strains used in this study and their plasmids.

	SA15303	SA14318	SA17155
Total size	5,144,932 bp	4,742,137 bp	5,240,120 bp
Plasmid size	286,565 bp	29,276 bp	199,024 bp
Plasmid type	IncHI2A	IncX1	IncHI2
Number of resistance genes on the plasmid	9	4	12
Number of homopolymer regions on the plasmid	131	7	94
MDR profiles	Chloramphenicol-Gentamicin-Tetracycline-Trimethoprim-Sulfamethoxazole	Aztreonam-Cefazolin-Ceftazidime-Ceftriaxone-Chloramphenicol-Ciprofloxacin-rtapenem-Meropenem-Imipenem-Tetracycline-Sulfamethoxazole	Cefazolin-Ceftazidime-Ceftriaxone-Gentamicin Chloramphenicol-Ciprofloxacin-rtapenem-Meropenem-Imipenem-Tetracycline-Sulfamethoxazole

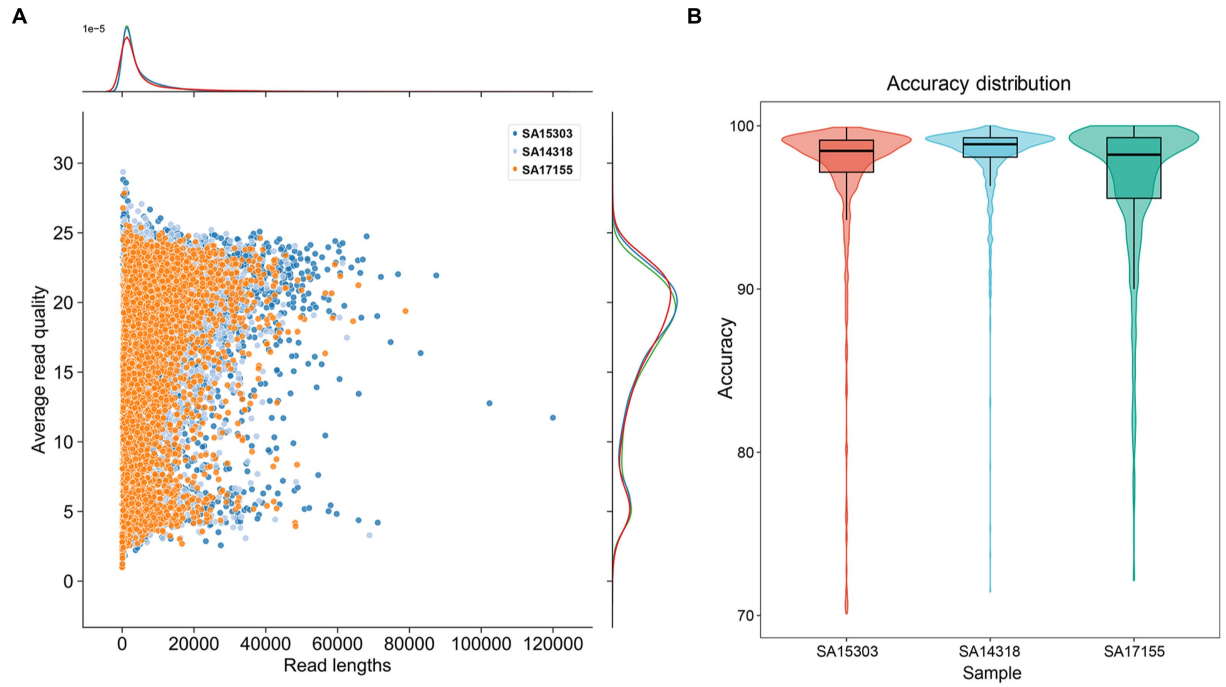


FIGURE 2
Summary of read quality. **(A)** The distribution of sequence length and sequence Q value of duplex data. **(B)** Box plot of read accuracy compared with reference data.

3.2. Genome assembly and error correction

After removing the low-quality reads, *de novo* assembly was performed using Flye to obtain preliminary results. Medaka was then run three times to correct errors. About 10× coverage (100 Megabyte, 25 min) was able to obtain complete plasmid sequences for all three strains. However, to obtain the complete sequence of both chromosomes and plasmids, coverage needs to be increased to 30× (300 Megabyte, approximately 75 min; Table 3).

Compared with the reference genome sequences, the errors generated can be classified as DELs, INS, and SNVs. SNVs represented the largest number of errors (48%), followed by DELs (26%). When the depth of the sequencing data reached 75× (750 Megabyte, 175 min), the assembly error rate was stable, and we obtained satisfactory sequences (Table 3). After assembly, the error-correction capabilities of Medaka were clearly visible; with DEL errors being the majority. Among them, 83.33% of DEL errors were eliminated, and, subsequently, 68.42% of SNV errors were corrected (Table 4). Medaka tends to fix DEL errors, which were more likely to occur in *de novo* assembly. The initial error correction impact of Medaka was notable, whereas the second and third error correction effects had no appreciable improvement (Supplementary Table 1).

10.3389/fmicb.2023.1179966

TABLE 2 Summary of sequencing data.

		SA15303	SA14318	SA17155
Sample concentration (ng/μL)		97.2	95.2	96.2
SQK-LSK114 kit and flow cell R10.4.1	Total file size (gigabytes)	5.03	4.73	4.95
	Depth	508×	497×	519×
	Total reads	361,153	309,825	318,461
	Reads > 5 kbp (%)	42.06	39.34	46.15
	Average length of reads (kbp)	5.83	4.98	5.18
	Percentage of Q10(%)	88.27	85.58	87.2
	Percentage of Q20(%)	52.67	42.31	46.25
	Average accuracy of reads (%)	98.94	98.88	98.83
	Alignment error rate (%)	2.19	2.11	1.99
SQK-LSK110 kit and flow cell R9.4	Total file size (gigabytes)	5.05	4.87	4.89
	Depth	512×	512×	511×
	Total reads	656,556	457,263	504,270
	Reads > 5kbp (%)	28.53	33.15	37.22
	Average length of reads (kbp)	3.96	4.59	4.14
	Percentage of Q10(%)	89.74	85.26	82.52
	Percentage of Q20(%)	13.65	14.22	18.09
	Average accuracy of reads (%)	94.02	92.53	93.84
	Alignment error rate (%)	5.79	7.18	5.11

TABLE 3 The number of errors in chromosome and plasmid sequences.

		Running time (min)	25	75	175	1,050
SA15303	Chromosome	Coverage depth	9×	31×	75×	477×
		Whether cyclic	N	Y	Y	Y
		Size (bp)	1,199,135	4,848,505	4,848,501	4,848,495
		Errors	532	10	9	5
	Plasmid	Coverage depth	11×	41×	94×	594×
		Whether cyclic	Y	Y	Y	Y
		Size (bp)	286,566	286,565	286,567	286,564
		Errors	15	2	2	3
SA14318	Chromosome	Coverage depth	13×	39×	81×	450×
		Whether cyclic	N	Y	Y	Y
		Size (bp)	3,517,587	4,712,839	4,712,855	4,712,849
		Errors	123	11	8	9
	Plasmid	Coverage depth	6×	22×	55×	306×
		Whether cyclic	Y	Y	Y	Y
		Size (bp)	29,281	29,266	29,276	29,276
		Errors	19	0	0	0
SA17155	Chromosome	Coverage depth	11×	31×	71×	466×
		Whether cyclic	N	Y	Y	Y
		Size (bp)	3,256,928	5,031,756	5,030,613	5,030,601
		Errors	535	12	9	8
	Plasmid	Coverage depth	9×	23×	51×	358×
		Whether cyclic	Y	Y	Y	Y
		Size (bp)	199,056	199,024	199,022	199,023
		Errors	50	4	2	3

10.3389/fmicb.2023.1179966

At a depth of sequencing data of 500×, the genomes of the three strains still had, respectively, 8, 9, and 11 errors that could not be corrected when compared with the reference genome sequences. Among SNV errors, G was frequently misidentified as T (8/9) and C was frequently misidentified as A (4/5). Duplication of the same base was the cause of insertion and deletion errors that could not be corrected. On chromosomes, 88.89% of the errors were found in homopolymer regions (Supplementary Table 2). Likewise, on plasmids, all INS and DEL errors were found in homopolymer regions (Figure 3). The reason for the above faults may be that Medaka is not adapted to the latest ONT model; thus, sequencing correction may improve with future software updates.

When the depth of sequencing data reached 75× coverage, the quality of the assembled sequences was near optimal. The average nucleotide identity was >99.9975% between each assembled sequence and the reference genome (Figure 4).

TABLE 4 Errors before and after correction.

	Type of variant	Before correction	After correction
SA15303	Variant-DEL	10	2
	Variant-INS	3	7
	Variant-SNP	3	2
SA14318	Variant-DEL	8	0
	Variant-INS	2	2
	Variant-SNP	7	5
SA17155	Variant-DEL	6	1
	Variant-INS	1	4
	Variant-SNP	9	6

Compared with sequences at 75× coverage depth, chromosomal sequences at 500× coverage depth had an average error reduction of 1.67, whereas plasmids had an average error increase of 0.67. Thus, although the 500× sequencing required a further 875 min and generated 6.5-fold more data than the 75× sequencing, the effects were not noticeably improved.

The latest ONT kit SQK-LSK114 and flow cell R10.4.1 have dramatically reduced errors in the assembled genomes. For data coverage that exceeded 75× (750 Megabyte, 160 min), using only Nanopore sequencing data in *de novo* genome packing, complete sequences were achieved that nearly matched the accuracy of NGS without requiring short-read data.

4. Discussion

The spread of ARGs carried by plasmids is a major public health issue (San Millan, 2018). As mobile genetic elements that can carry ARGs and be transferred easily between different bacterial species, plasmids enable the quick and effective dissemination of ARGs (Vit et al., 2020). The emergence of MDR and the spread of drug resistance between bacterial strains can both be aided by plasmids (Paterson and van Duin, 2017). It is critical to comprehend how bacteria acquire and spread resistance to antibiotics, in addition to the molecular mechanisms of this phenomenon (Jordt et al., 2020). High-quality finished genomes of drug-resistant bacteria are required to monitor the transmission of antimicrobial resistance. The current tools for ARG detection are known to be highly accurate when used with short-read sequencing data. However, short-read sequencing cannot provide accurate localization of ARGs, which can be carried by both bacterial plasmids and chromosomes. More frequent cross-host resistance epidemic events initiated by plasmids carrying ARGs have occurred (Li et al., 2022), suggesting the greater potential for ARG transmission through plasmids in humans, food, animals, and the environment. To

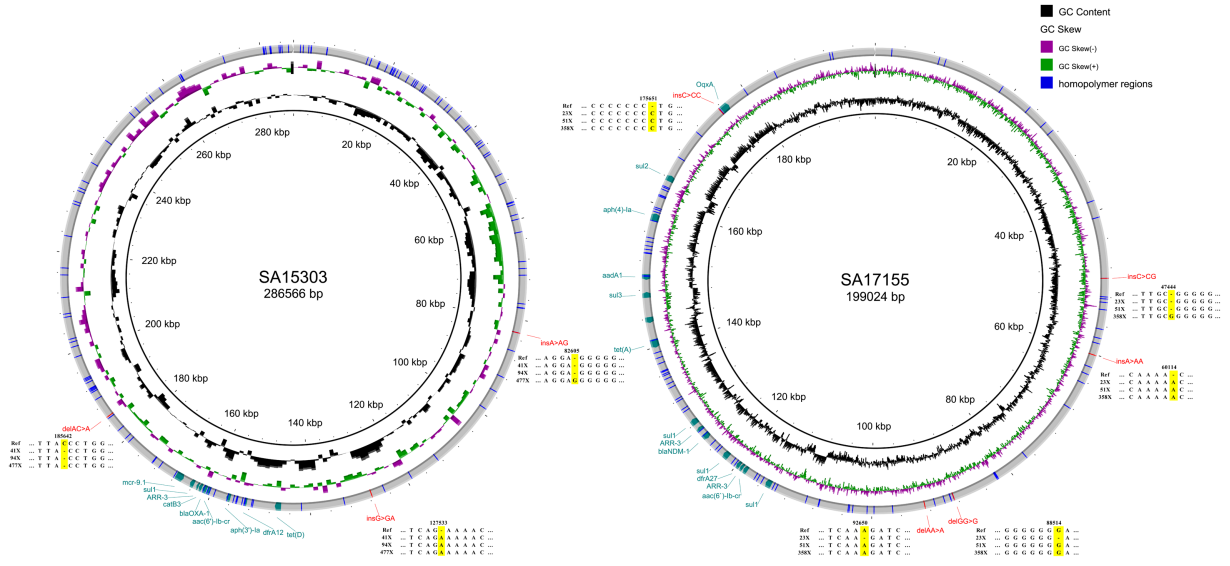


FIGURE 3
Plasmid structure diagrams for strains SA15303 and SA17155. Compared with the reference sequences, the errors at 30× coverage depth were showed. The plasmid of SA14318 has no errors and was not shown in this figure.

10.3389/fmicb.2023.1179966

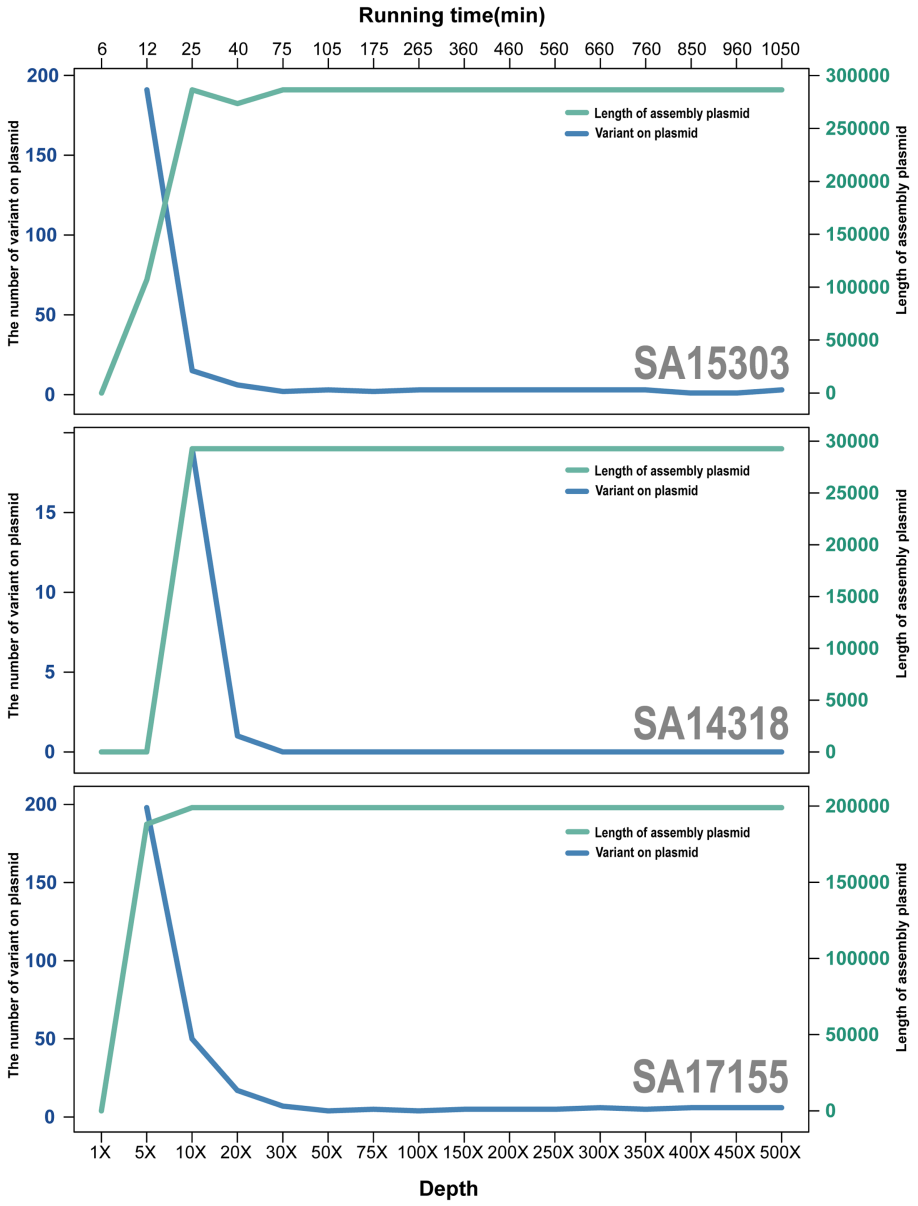


FIGURE 4
Trend plot of plasmid size and errors with increasing sequencing depth. Green lines and axes represent the size of plasmids, and blue lines represent the number of errors.

obtain high-quality finished genomes, it has been necessary to include short-read polishing of long-read sequencing data to correct errors (Senol Cali et al., 2019). This combined approach required two sequencing libraries and was time-consuming, difficult to perform, and can entail high costs. The optimal solution is to improve the accuracy of long-read sequencing. Based on long-read sequencing, one can locate ARGs on chromosomes or plasmids. Horizontal transfer of plasmids plays an important role in the spread of multidrug-resistant bacteria, identifying plasmid-borne resistance genes is necessary to estimate the spread of resistance among bacteria. The coexistence of ARGs is very common, and poses significant public health and food safety threats. Obtaining complete plasmid sequences is the most effective means of detecting ARGs and the coexistence of ARGs on the same plasmid.

In this study, the latest ONT SQK-LSK114 kit with flow cell R10.4.1 was used, and the results demonstrated that a high-quality finished genome could be obtained using *de novo* assembly. Compared to the previous process, the Q-score of reads was significantly improved (Smith et al., 2020). In previous studies, the limitation of ONT sequencing was the relatively high error rate, which in some cases can reach 10% (Khrenova et al., 2022). Low base-call accuracy has limited the reliability of ONT data for critical genomic epidemiology tasks such as ARG and virulence gene detection and typing, serotype prediction, and cluster identification (Foster-Nyarko et al., 2023). In this study, the average read accuracy was 98.9%, indicating that the novel approach has the potential to greatly improve read accuracy. When using the latest ONT SQK-LSK114 kit and R10.4.1 flow cell for long-read sequencing, error correction using

short-read sequence data is not required. The novel ONT sequencing method saves money and time. Both the preparation of short- and long-read libraries as well as sequencing, which take at least 1 day, are unnecessary.

Regarding the new ONT sequencing method, once low-quality reads had been removed, the accuracy of the assembled sequences was nearly identical to that based on NGS. However, homopolymer regions increased the possibility of INS and DEL errors during assembly, and biased SNV errors still needed to be improved. In the future, the analysis pipeline may become more streamlined and effective.

In conclusion, use of the ONT LSK114 kit combined with flow cell R10.4.1 improved the accuracy of *de novo* genome assembly. When sequencing bacteria and plasmid, high-quality complete genomes with ideal coverage and identity were obtained without short-read or reference polishing. To obtain saturated raw data, we recommend acquiring a depth of 100× raw data (1 Gigabytes, 265 min), and performing error correction three times after assembly. This method will save time in obtaining high-quality complete genomes of bacteria for antimicrobial resistance surveillance, and is likely to become a valuable tool for monitoring the transmission of plasmid-borne drug resistance genes (Peter et al., 2020).

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

XL, ZL, and BK conceived and designed this study. BP, ML, YP, and JX contributed to the experiment. WZh and WZe contributed to writing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (2020YFE0205700 and 2022YFC2303900), the Major Projects of the National Natural Science Foundation of China (22193064), and the Science Foundation (2022SKLID303) of the State Key Laboratory of Infectious Disease Prevention and Control, 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.2023.1179966/full#supplementary-material

References

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., 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

Aslam, B., Khurshid, M., Arshad, M. I., Muzammil, S., Rasool, M., Yasmeen, N., et al. (2021). 'Antibiotic resistance: one health one world outlook. *Front. Cell. Infect. Microbiol.* 11:771510. doi: 10.3389/fcimb.2021.771510

Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., and Scorilas, A. (2021). Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. *Life* 12:30. doi: 10.3390/life12010030

Bengtsson-Palme, J., Kristiansson, E., and Larsson, D. G. J. (2018). Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiol. Rev.* 42:fux053. doi: 10.1093/femsre/fux053

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria: plasmid-encoded antibiotic resistance. *Br. J. Pharmacol.* 153, S347–S357. doi: 10.1038/sj.bjp.0707607

Boostrom, I., Portal, E. A. R., Spiller, O. B., Walsh, T. R., and Sands, K. (2022). Comparing long-read assemblers to explore the potential of a sustainable low-cost, low-infrastructure approach to sequence antimicrobial resistant Bacteria with Oxford Nanopore sequencing. *Front. Microbiol.* 13:796465. doi: 10.3389/fmicb.2022.796465

Cohen, K. A., Manson, A. L., Desjardins, C. A., Abeel, T., and Earl, A. M. (2019). Deciphering drug resistance in *Mycobacterium tuberculosis* using whole-genome sequencing: progress, promise, and challenges. *Genome Med.* 11:45. doi: 10.1186/s13073-019-0660-8

de Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

Ding, Y., Saw, W. Y., Tan, L. W. L., Moong, D. K. N., Nagarajan, N., Teo, Y. Y., et al. (2020). Emergence of tigecycline- and eravacycline-resistant Tet(X4)-producing Enterobacteriaceae in the gut microbiota of healthy Singaporeans. *J. Antimicrob. Chemother.* 75, 3480–3484. doi: 10.1093/jac/dkaa372

Doyle, D., Peirano, G., Lascols, C., Lloyd, T., Church, D. L., and Pitout, J. D. D. (2012). Laboratory detection of Enterobacteriaceae that produce carbapenemases. *J. Clin. Microbiol.* 50, 3877–3880. doi: 10.1128/JCM.02117-12

Fan, J., Cai, H., Fang, Y., He, J., Zhang, L., Xu, Q., et al. (2022). Molecular genetic characteristics of plasmid-borne mcr-9 in *Salmonella enterica* serotype Typhimurium and Thompson in Zhejiang, China. *Front. Microbiol.* 13:852434. doi: 10.3389/fmicb.2022.852434

Foster-Nyarko, E., Cottingham, H., Wick, R. R., Judd, L. M., Lam, M. M. C., Wyres, K. L., et al. (2023). Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. *Microb Genomics* 9:mgen000936. doi: 10.1099/mgen.0.000936

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

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

Jordt, H., Stalder, T., Kosterlitz, O., Ponciano, J. M., Top, E. M., and Kerr, B. (2020). Coevolution of host–plasmid pairs facilitates the emergence of novel multidrug resistance. *Nat Ecol Evol* 4, 863–869. doi: 10.1038/s41559-020-1170-1

Kathirvel, K., Rudhra, O., Rajapandian, S. G. K., Venkatesh Prajna, N., Lalitha, P., and Devarajan, B. (2021). Characterization of antibiotic resistance and virulence genes of

ocular methicillin-resistant *Staphylococcus aureus* strains through complete genome analysis. *Exp. Eye Res.* 212:108764. doi: 10.1016/j.exer.2021.108764

Khrenova, M. G., Panova, T. V., Rodin, V. A., Kryakvin, M. A., Lukyanov, D. A., Osterman, I. A., et al. (2022). Nanopore sequencing for De novo bacterial genome assembly and search for single-nucleotide polymorphism. *Int. J. Mol. Sci.* 23:8569. doi: 10.3390/ijms23158569

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. doi: 10.1038/s41587-019-0072-8

Li, Z., Li, Z., Peng, Y., Lu, X., and Kan, B.State Key Laboratory of Infectious Disease Prevention and Control; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China (2022). Trans-regional and cross-host spread of mcr-carrying plasmids revealed by complete plasmid sequences — 44 countries, 1998–2020. *China CDC Weekly* 4, 242–248. doi: 10.46234/ccdcw2022.058

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14, 265–279. doi: 10.1016/j.gpb.2016.05.004

Malhotra-Kumar, S., Xavier, B. B., das, A. J., Lammens, C., Butaye, P., and Goossens, H. (2016). Colistin resistance gene mcr-1 harboured on a multidrug resistant plasmid. *Lancet Infect. Dis.* 16, 283–284. doi: 10.1016/S1473-3099(16)00012-8

Nathan, C. (2020). Resisting antimicrobial resistance. *Nat. Rev. Microbiol.* 18, 259–260. doi: 10.1038/s41579-020-0348-5

Nikaido, H. (2009). Multidrug Resistance in Bacteria. *Annu. Rev. Biochem.* 78, 119–146. doi: 10.1146/annurev.biochem.78.082907.145923

Partridge, S. R., Tsafnat, G., Coiera, E., and Iredell, J. R. (2009). Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol. Rev.* 33, 757–784. doi: 10.1111/j.1574-6976.2009.00175.x

Paterson, D. L., and van Duin, D. (2017). 'China's antibiotic resistance problems', the lancet. *Infect. Dis.* 17, 351–352. doi: 10.1016/S1473-3099(17)30053-1

Peter, S., Bosio, M., Gross, C., Bezdan, D., Gutierrez, J., Oberhettinger, P., et al. (2020). Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by Nanopore sequencing. *mSphere* 5, e00525–e00520. doi: 10.1128/mSphere.00525-20

Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., and Tsongalis, G. J. (2019). Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of Nanopore sequencing. *J. Clin. Microbiol.* 58, e01315–e01319. doi: 10.1128/JCM.01315-19

San Millan, A. (2018). Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol.* 26, 978–985. doi: 10.1016/j.tim.2018.06.007

Senol Cali, D., Kim, J. S., Ghose, S., Alkan, C., and Mutlu, O. (2019). Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief. Bioinform.* 20, 1542–1559. doi: 10.1093/bib/bby017

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., et al. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* 19, 823–826. doi: 10.1038/s41592-022-01539-7

Smith, C., Halse, T. A., Shea, J., Modestil, H., Fowler, R. C., Musser, K. A., et al. (2020). Assessing Nanopore sequencing for clinical diagnostics: a comparison of next-generation sequencing (NGS) methods for *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 59, e00583–e00520. doi: 10.1128/JCM.00583-20

Vit, C., Loot, C., Escudero, J. A., Nivina, A., and Mazel, D. (2020). Integron identification in bacterial genomes and cassette recombination assays. *Methods Mol. Biol.* 2075, 189–208. doi: 10.1007/978-1-4939-9877-7_14

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

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., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* 20:129. doi: 10.1186/s13059-019-1727-y

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 May 2023
DOI 10.3389/fmicb.2023.1165916

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Álvaro Chiner-Oms,
Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO), Spain
Julius Mugweru,
University of Embu, Kenya

*CORRESPONDENCE
Tao Luo
✉ taoluo.scu@foxmail.com
Hong-Sheng Wang
✉ whs33@vip.sina.com

RECEIVED 14 February 2023
ACCEPTED 19 April 2023
PUBLISHED 17 May 2023

CITATION
Mei Y-M, Zhang W-Y, Sun J-Y, Jiang H-Q, Shi Y, Xiong J-S, Wang L, Chen Y-Q, Long S-Y, Pan C, Luo T and Wang H-S (2023) Genomic characteristics of *Mycobacterium tuberculosis* isolates of cutaneous tuberculosis. *Front. Microbiol.* 14:1165916. doi: 10.3389/fmicb.2023.1165916

COPYRIGHT
© 2023 Mei, Zhang, Sun, Jiang, Shi, Xiong, Wang, Chen, Long, Pan, Luo and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Genomic characteristics of *Mycobacterium tuberculosis* isolates of cutaneous tuberculosis

You-Ming Mei[1,2], Wen-Yue Zhang[1,2], Ji-Ya Sun[3,4], Hai-Qin Jiang[1,2], Ying Shi[1,2], Jing-Shu Xiong[1,2], Le Wang[1,2], Yan-Qing Chen[1,2], Si-Yu Long[1,2], Chun Pan[1,2], Tao Luo[5]* and Hong-Sheng Wang[1,2]*

[1]Hospital of Skin Disease, Institute of Dermatology, Chinese Academy of Medical Sciences and Peking Union Medical College, Nanjing, China, [2]Jiangsu Key Laboratory of Molecular Biology for Skin Diseases and STIs, Nanjing, China, [3]Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou, China, [4]Suzhou Institute of Systems Medicine, Suzhou, China, [5]Department of Pathogen Biology, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, China

Objectives: Cutaneous tuberculosis with various manifestations can be divided into several clinical types according to the host's immune status and infective route. However, the etiological factors of this disease remain unclear. The objective of this study is to investigate the pathogens associated with the occurrence and different types of cutaneous tuberculosis.

Methods: 58 *Mycobacterium tuberculosis* strains isolated from cutaneous tuberculosis over the last 20 years were sequenced and analyzed for genomic characteristics including lineage distribution, drug-resistance mutations, and mutations potentially associated with different sites of infection.

Results: The *M. tuberculosis* strains from four major types of cutaneous tuberculosis and pulmonary tuberculosis shared similar genotypes and genomic composition. The strains isolated from cutaneous tuberculosis had a lower rate of drug resistance. Phylogenic analysis showed cutaneous tuberculosis and pulmonary tuberculosis isolates scattered on the three. Several SNPs in metabolism related genes exhibited a strong correlation with different infection sites.

Conclusions: The different infection sites of TB may barely be affected by large genomic changes in *M. tuberculosis* isolates, but the significant difference in SNPs of drug resistance gene and metabolism-related genes still deserves more attention.

KEYWORDS

mycobacteria infection, *Mycobacterium tuberculosis*, whole genome sequence, cutaneous tuberculosis, comparative genomics

Introduction

The *Mycobacterium tuberculosis* complex consists of several groups with geographic origins and distributions. Among them, *M. tuberculosis sensu stricto* (MTB, Lineages 1, 2, 3, 4, and 7) and *M. tuberculosis var. africanum* (Lineages 5 and 6) are the major pathogens of human disease. Tuberculosis, which is caused by infection with the *M. tuberculosis* complex, remains a threat to global public health (World Health Organization, 2021). It can be simply divided into pulmonary tuberculosis (PTB) and extrapulmonary tuberculosis (EPTB). EPTB is less contagious and has received secondary attention in most tuberculosis surveillance programs, but research has revealed an increased proportion of this disease globally (Sandgren et al., 2013; Gomes et al., 2014). Cutaneous tuberculosis (CTB) is one of the rare forms of EPTB, and it accounts for only 1–2% of tuberculosis cases (Franco-Paredes et al., 2018). CTB is conventionally classified into several groups with

characteristic manifestations. The various clinical forms were previously supposed to be the result of both different infective routes and host immune statuses. Lupus vulgaris (LV), tuberculosis cutis orificialis (TCO), and scrofuloderma (SFD) are considered to be disseminators of visceral tuberculosis, while tuberculosis verrucosa cutis (TVC) and tuberculosis chancre are supposed to develop through exogenous inoculation (Hill and Sanders, 2017). However, the current classification rule is mainly derived from clinical inference. The temperature is an important factor for the growth of the microorganism. Mycobacteria such as *M. marinum* show comparatively strict temperature limits in human infection. In general, the skin provides a more stressful environment than other visceral organs, including the lower temperature and hypoxia, which are unfit for MTB infection and proliferation, which may suggest the special fitness of these CTB isolates (Jabir et al., 2017). However, we found few studies focusing on the potential etiological factors of CTB or reporting the characteristics of CTB isolates.

Approaches based on whole genome sequencing provide superior insights into mutation-based mycobacterial genotyping, drug resistance profiling, and transmission detection. Analysis models based on comparative genomics have successfully identified several novel resistance or adaptation-related mutations in MTB (Desjardins et al., 2016; Farhat et al., 2019). Many studies discuss MTB isolates from pulmonary sites or other extrapulmonary sites, but the genome sequence of the CTB isolate has not been reported to date (Sandgren et al., 2013; Gomes et al., 2014; Desjardins et al., 2016). Thus, this study collected CTB isolates in China to investigate the pathogen genomic characteristics of CTB.

Methods

Sample collection

We selected 111 cases suspected as CTB confirmed by PCR amplification and sequencing for *16S*, *rpoB*, or *hsp65* genes of skin biopsy, or by sequencing of tissue cultures in the Hospital of Dermatology, Chinese Academy of Medical Sciences from 2000 to 2020. All the clinical records for these cases were retrospectively reviewed. The patients' skin biopsy was collected before anti-TB therapy and ground and spread on improved Lowenstein–Jensen slants at 32 and 37°C, respectively. The clinical culture time till visible colonies appeared in 87 strains was recorded (Table 1). The remaining organisms after identification were routinely freeze-preserved at −80°C. In 2021, all the existing frozen stocks (73) were sub-cultured at 37°C for 4–12 weeks to acquire enough organisms for sequencing. The stored strains unsuccessfully recovered or had contamination were excluded in the following study. Finally, 58 MTB strains isolated from CTB lesions were included in this study.

Whole genome sequencing and SNP calling

The organisms were lysed with 20 mg/ml lysozyme for cell wall disruption, and then, DNA was extracted using a QIAamp DNA Blood Mini Kit (Qiagen, Manchester, UK). DNA libraries with lengths of 150 bp were generated and sequenced on an Illumina HiSeq 4000 platform (Illumina, San Diego, CA, USA) at the Beijing Genomics Institute. The sequenced reads were assembled using SOAPdenovo with the reference of MTB H37Rv (Luo et al., 2012). Four strains from different CTB clinical types were sequenced using PacBio RS II and an Illumina HiSeq 4000 platform. PacBio subreads with a length of < 1 kb were removed. Draft unitigs were assembled using Celera Assembler. The non-coding RNA was detected using tRNAscan-SE, RNAmmer, and the Rfam database (Schattner et al., 2005; Karin et al., 2007; Kalvari et al., 2018). Prophage regions were predicted using PhiSpy, and CRISPR was detected by CRISPRCasFinder (Akhter et al., 2012; Couvin et al., 2018). Rapid Annotation using Subsystem Technology (RAST) was used for gene annotation (Aziz et al., 2008). All sequence data associated with this study are deposited in the NCBI Bioproject, accession PRJNA820632.

In searching for whole genome sequencing samples of MTB from southeastern China in NCBI and EMBL databases, we found 219 strains from Shanghai in the same period (PRJNA417753) (Yang C. et al., 2018). The adapters and raw reads of CTB and PTB isolates with low quality (>40% bases with base quality of < 20 or N base of >5) were trimmed using Fastp and were checked by FastQC and MultiQC for read quality (Ewels et al., 2016; Chen et al., 2018). The data sets were aligned to the reference genome sequence of *Mycobacterium tuberculosis* H37Rv (Genbank accession: GCF_000195955.2) using BWA mem (Li and Durbin, 2009). Duplicated reads were removed by Sambamba (Tarasov et al., 2015). Genomes with a read depth of <50× or a coverage of <95% were excluded. Finally, 188 PTB isolates collected in 2008–2015 were selected as the group of PTB isolates.

Bioinformatic analysis

The molecular drug resistance to TB drugs and the lineage distribution of the genomes were examined on TB-profiler trimmed data sets (Phelan et al., 2019). Roary was used to create a pan-genome, and Scoary was used to analyze the accessory gene presence or absence in CTB isolates with TCO or TVC (Page et al., 2015; Brynildsrud et al., 2016). Orthovenn2 was used for the comparative analysis of whole-genome orthologous clusters of the four genomes, representing the four CTB types (Xu et al., 2019). Descriptive statistics were used to describe the characteristics of the isolates. Chi-square tests and 95% confidence intervals (CIs) were used to compare non-metrological data by SPSS version 25 (SPSS Inc., Illinois, USA). For all analyses, a p-value of <0.05 was considered to be statistically significant.

Genome-wide association and convergence test

For the phylogeny construction and convergence test, variant calling was performed on the whole genome alignment of all the CTB and PTB strains by Snippy (Seemann, 2023). A total of 8,271 highly credible SNPs used for phylogeny construction were obtained after excluding the regions of the drug target gene, PE/PPE family, and genes with missing sites in >10% of samples. The phylogenetic tree was reconstructed using IQ-TREE

10.3389/fmicb.2023.1165916

TABLE 1 Clinical information of cutaneous tuberculosis and the growth temperature of these isolates.

		LV	TVC	TCO	SFD	Total
Case		56 (64.37)	15 (17.24%)	11 (12.64)	5 (5.75%)	87
Gender (F/M)		24/32	12/3	11/0	1/4	62/49
Age		20–84	6–86	38–71	51–77	6–86
Mean age		53.09	52.57	55.00	67.00	53.46
PTB history		5 (8.93%)	0 (0%)	3 (27.27%)	1 (20%)	9 (10.34%)
Growth temperature	32°C	6 (10.71%)	3 (20%)	3 (27.27%)	0 (0%)	12 (13.79%)
	37°C	35 (62.5%)	6 (40%)	7 (63.64%)	4 (80%)	52 (59.77%)
	32/37°C	3 (5.36%)	2 (13.33%)	1 (9.09%)	0 (0%)	6 (6.9%)
	NA	12	4	0	1	17
Drug-resistant mutation		3 (5.36%)	2 (13.33%)	0 (0%)	0 (0%)	5 (5.75%)
Lineage	Lineage 2	26 (46.43%)	7 (46.67%)	4 (36.36%)	2 (40)%	39 (44.83%)
	Lineage 4	8 (14.29%)	3 (20%)	2 (18.18%)	0 (0%)	13 (14.94%)

TABLE 2 Drug resistance mutations in cutaneous tuberculosis isolates.

Drug	Mutation	Strains
RFP	rpoB_p.Ser450Leu	1
INH	katG_p.Ser315Thr	4
EMB	embB_p.Met306Ile	1
STR	rpsL_p.Lys88Arg	3
	gid_c.115_115del	1
PAS	thyA_p.His75Asn	2
DR type	**Drug**	**Strains**
Drug resistant	INH	1
	STR	1
	PAS	2
	INH + STR	2
MDR	RFP + INH + EMB + STR	1

assuming the GTR+F model with 1,000 bootstrap replicates, and annotation and visualization were processed by iTOL (Letunic and Bork, 2011; Nguyen et al., 2015). The convergence test was performed using the concatenated alignment of the SNP site of the 58 CTB isolates and the tree above (Page et al., 2016). The homoplastic mutations independently occurring in different clades were found in the particular site with the top frequency of change along the tree with tree-time (Sagulenko et al., 2018). GWAS analyses were performed on non-synonymous SNPs in CDS using the burden test in PYSEER (Lees et al., 2018). Pairwise distances calculated from the phylogeny tree were used to account for relatedness. A *P*-value threshold of <1E-04 was considered to be statistically significant. The products and the likely impact of the gene with SNP were annotated by SnpEff (Cingolani et al., 2012). The published algorithm was used to predict the effects of identified mutations on protein stability (Capriotti et al., 2005).

Results

Clinical information and laboratory features of CTB isolates

A total of 111 HIV-negative cases with a positive culture and conserved gene sequencing results of MTB from 2000 to 2020 were reviewed in this study. Among the cases with intact geographic information, most of them came from Jiangsu (53.47%, 54/101) and Anhui provinces (26.73%, 27/101) (Table 1). The average age of all patients was 53.46 ± 16.26 years (range from 6 to 86 years, median of 54), and 55.86% (62/111) were male patients. A total of 87 cases could be categorized into four major CTB classifications based on medical records, images, and pathological examination results, and nine patients (10.34%) had signs of active infection or a history of visceral tuberculosis. LV was the major form of CTB (64.37%, 56/87), followed by TVC (15/87, 17.24%) and TCO (11/87, 12.64%), which is similar to previous literature reports in China (Zhang et al., 2018). Among the 45 patients with intact treatment records, 93.33% (42/45) of patients recovered well after 6 months of standard multidrug therapy. Three cases recurred after anti-tuberculosis therapy more than 10 years later.

Genomic characteristics of CTB isolates

A total of 58 CTB isolates were successfully recovered, were sequenced on the Illumina HiSeq 4000 platform, and were assembled reads using SOAPdenovo with the reference of MTB H37Rv (Methods). After all the CTB isolates were examined on TB-profiler trimmed reads, these CTB isolates were confirmed as MTB with no isolates of other subspecies, such as *M. bovis*, found. Lineage 2.2 (Beijing family strains) predominated in all cases (41/58, 70.69%). The remaining 17 isolates belonged to Lineage 4, in which Lineage 4.4 was the most prevalent sub-lineage (8/17, 47.06%), followed by Lineages 4.2

TABLE 3 Drug resistance strain distribution in isolates of cutaneous tuberculosis and pulmonary tuberculosis in adjacent provinces.

	CTB	PTB	
DR	$n = 58$ (Jiangsu, 2000–2019)	$n = 260$, (Jiangsu, 2010) (Liu et al., 2011)	$n = 2,133$, (Anhui, 2015–2016) (Yao et al., 2020)
RFP	1 (1.72%)	38 (14.62%)	163 (7.64%)
INH	4 (6.90%)	45 (17.31%)	223 (10.45%)
EMB	1 (1.72%)	20 (7.69%)	40 (1.88%)
STR	4 (6.90%)	43 (16.54%)	234 (10.97%)
Mono-DR	4 (6.90%)	33 (12.7%)	-
MDR	1 (1.72%)	34 (13.1%)	106 (4.97%)
XDR	-	-	13 (0.61%)
Total	7 (12.07%)	74 (28.5%)	-

TABLE 4 Genomic information of the four cutaneous tuberculosis isolates from different clinical types.

	SGF0232017	SGF0422017	SGF0472017	SGF0702019	*M. tuberculosis* H37Rv
Clinical type	TVC	LV	SGD	TVC	
Length	4,413,076	4,402,590	4,417,237	4,419,329	4,411,532
GC (%)	65.60	65.61	65.61	65.62	65.6
Coding sequences	4,298	4,286	4,330	4,303	4,299
Subsystems	293	293	293	292	294
Number of prophage	5	6	10	5	5
Number of CRISPR	8	8	8	9	9
tRNA (% in genome)	45	45	45	45	45
5s_rRNA (% in genome)	1	1	1	1	1
16s_rRNA (% in genome)	1	1	1	1	1
23s_rRNA (% in genome)	1	1	1	1	1
sRNA (% in genome)	15	15	15	15	15

and 4.5 (4/17, 23.53%, respectively). No strains of Lineage 1 or Lineage 3 were found in these specimens. The lineage, growth time, or temperature of strains of different CTB types revealed no significant differences. Seven CTB strains were found that had resistance mutations, and they accounted for only 12.07% of all the CTB isolates (Tables 2, 3). Most of these isolates were mono-drug resistant (4/7, 57.14%), and they were mainly resistant to isoniazid or streptomycin. The only MDR isolate detected with resistance to rifampicin, isoniazid, ethambutol, and streptomycin was collected from a recurrent LV after 20 years of treatment. A higher proportion of drug resistance was observed in Lineage 2.2 strains (14.63%, 6/41) than in Lineage 4 strains (5.88%, 1/17), with no significant difference ($\chi^2 = 0.239$, $p = 0.625$).

The covered length of the draft CTB genomes ranges from 4,348,399 to 4,406,327 bp (98.57–99.88%) compared with the reference MTB H37Rv (Supplementary Table 1). These genomes have a GC content between 64.10 and 65.61%. In the pan-genome analysis, 5,219 orthologous genes including 3,518 core genes and other variable components (Softcore: 247, Cloud: 1,032, Shell: 422) were identified. The pan-genome curve indicated an open pan-genome of MTB as previously reported (Yang C. et al., 2018). In a comparison of TVC and TCO, which were the specific CTB types caused by exogenous and endogenous infections, respectively, we found that the existence of several genes significantly differed in isolates of TVC and TCO, including *PPE47* (6/6 in TVC, 4/10 in TCO), *embR* (3/6 in TVC, 10/10 in TCO), and *moaA1* (3/6 in TVC, 10/10 in TCO), as well as three coding genes of putative proteins. However, these differences need to be further confirmed because the technical errors in short-read sequencing may not be totally eliminated. We performed PacBio genome sequencing in four representative Lineage 2 strains to acquire more precise *de novo* assemblies of different CTB groups to thoroughly investigate the genome of isolates from different CTB types. The basic information of the four assemblies is summarized in Table 4. A total of 4,286–4,303 protein-coding genes and 59 non-coding RNAs including 45 tRNAs, 3 rRNAs, and 15 sRNAs were annotated. The orthologous gene clusters showed similar compositions of these strains, and only the isolate of SFD was annotated with an additional unique gene of the transposase of the IS1081 element (Figure 1).

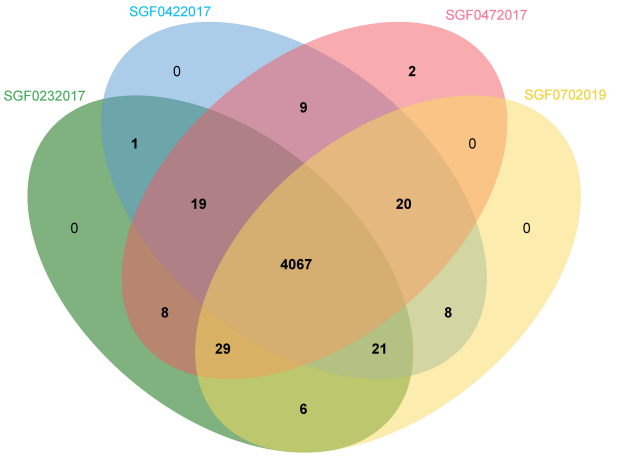


FIGURE 1
Venn image of the orthologous genes of four representative genomes from different CTB types.

Comparative genomic analysis of MTB isolates from CTB and PTB

We selected whole genome sequences of 188 PTB isolates from Shanghai from 2009 to 2015 to compare the MTB isolates from different infection sites. These strains had a significantly higher proportion of Lineage 2.2 (177/188, 94.15%) compared to CTB strains. The drug resistance mutation rate in the PTB group (48/188, 25.53%) was significantly higher compared to the CTB group ($\chi^2 = 4.286$, $p = 0.031$, OR = 2.498, 95%CI: 1.062–5.876) (Yang T. et al., 2018). To eliminate selection bias caused by the potential geographic influence due to the genetic similarity of PTB isolates in this city or by our filter process for low-quality PTB samples, we further compared the lineage distribution and drug resistance rate of CTB strains with PTB strains from Jiangsu or Anhui province in the same period. The genotype distribution showed a similar proportion of Lineage 2.2 (Beijing family), but a much lower drug resistance mutation rate could still be found in CTB strains (Liu et al., 2011; Yao et al., 2020) (Table 3).

In all the CTB isolates, 2,197 SNPs including 1,168 (60.77%) non-synonymous mutations, 734 (38.35%) synonymous SNPs, and 1,738 INDELs (751 insertions and 987 deletions) were detected. The phylogenetic tree constructed on whole genome SNPs between both PTB and CTB isolates showed that the CTB isolates were scattered on the tree. No clade formed by any specific clinical group was observed (Figure 2). SNP diversity among CTB strains ranged from 114 to 1,263, with an average of 654 SNPs. This finding revealed a high degree of genetic heterogeneity, and no transmission incident was detected among the CTB isolates in the study period.

Genome-wide association study (GWAS) is an effective tool in identifying genetic variants related to different clinical forms. Given the small dataset of CTB isolates and the existence of many low-frequency variants, a gene-based burden test was used to find the cumulative effect of these mutations. Out of 178 genes with non-synonymous variants correlated with the two infection forms ($p < 10^{-2}$), the following four genes were identified to have strong associations ($p < 10^{-4}$): Rv0392c (10.34% of CTB, 15.96% of PTB), Rv2088 (63.79% of CTB, 91.49% of PTB), Rv2331 (0.00% of CTB, 4.26% of PTB), and Rv3829c (5.17% of CTB, 14.89% of PTB). In the homoplastic analysis of the CTB group, we found two missense variations that were supposed to be homoplastic SNPs. They were also detected in PTB strains, namely, Rv1872c (c.757G>A), the coding gene of *lldD2*, had variation in 72.41% of CTB (42/58) and 95.21% of PTB (179/188) and Rv1873 (c.8C>T) was in the subsequent region of Rv1872c, was found in 3.45% of CTB (2/58), and was not detected in PTB.

Discussion

Among the current nine lineages of the *M. tuberculosis* complex, the Beijing family (Lineage 2.2) and Euro-American lineages (Lineage 4) are the most prevalent lineages in China (Napier et al., 2020). The CTB isolates in this study were also composed of strains of Lineage 2.2 and Lineage 4, which conform to the MTB epidemic status in China. The Beijing family was predominant in both PTB and CTB strains, while Lineage 4 had diverse genotypes in CTB isolates. Lineage 4 highly prevails in the western regions of China. Thus, we suggest that the highly diverse Lineage 4 subtypes in CTB isolates may be caused by the CTB patients coming from multiple regions. The included strains showed scattered distribution in each clade, and no cluster of the same clinical form was found on the phylogenetic tree, which suggests that the isolates of different infection sites were in similar evolutionary positions (Figure 1). Similarly, no significant difference in the accessory genome was found among different CTB types. SMRT sequencing also showed the coincident genomic composition of the four representative isolates and the MTB-type strain. Therefore, we consider that the clinical diversity of CTB has little relationship with the genetic changes in MTB. The lower consistency of CTB isolates is probably associated with the sporadic incidence of CTB (Lin et al., 2020).

Unlike PTB, CTB is generally neglected in many countries. The reason is that CTB is less contagious than PTB. In addition, CTB also usually responds well to first-line anti-tuberculosis therapy (Hill and Sanders, 2017; Zhang et al., 2018). In this study, the drug target mutation rate of CTB isolates was significantly lower than that of PTB isolates, which coincides with the high cure rates of CTB (93.33%). In the comparison of the drug-resistant mutant proportion with PTB isolates from Jiangsu and Anhui provinces, the significantly lower drug-resistant mutant rate showed no relationship with geographical distribution (Liu et al., 2011; Yao et al., 2020). CTB often presents as a chronic infection with mild symptoms, and MTB has few opportunities to spread MTB among patients. Thus, we suggest that the less drug-resistant mutation of these isolates may be due to a long infection time and less transmission or drug treatment experience (Allué-Guardia et al., 2021).

We used a genome-wide association study based on both homoplasy counting and allele counting approaches to find any subtle genetic variations that may cause different phenotypes. Two missense mutations in CTB isolates were detected as

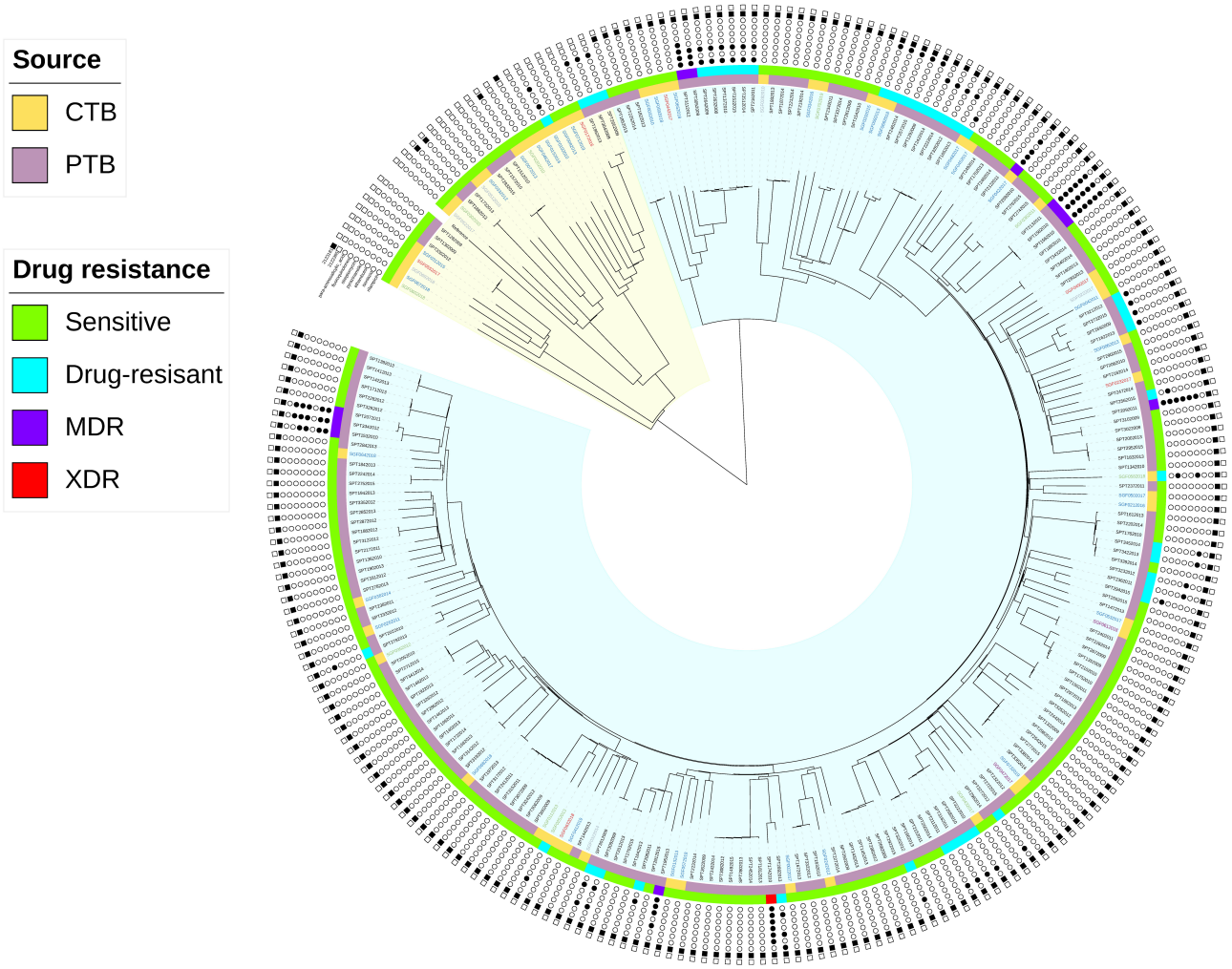


FIGURE 2
Phylogeography of the 246 MTB isolates (58 CTB and 188 PTB). Strains contained the SNPs on drug targets (INH, RFP, EMB, PNZ, STR, FLQ, and PASA) and loci under convergent evolutionary (Rv1872c (c.757G>A), 2123281; Rv1873 (c.8C>T), 2122395) are marked with a black dot. The branches in the yellow and blue range of the tree belong to Lineage 4 and Lineage 2, respectively. The inner stripes in orange (CTB) and purple (PTB) show the infective location, and the outer stripes in green (sensitive), blue (DR), purple (MDR), and red (XDR) show the drug sensitivity of the strains. MDR was defined as MTB strains with resistance to rifampicin and isoniazid. XDR was defined as MDR strains with resistance to any fluoroquinolone (FQ) and any second-line injectable drug.

homoplastic SNPs. Rv1872c, which is the coding gene of L-lactate dehydrogenase (*lldD2*), is essential for L-lactate reversion as a carbon source *in vitro* (Billig et al., 2017). The knock-out variant failed to utilize L-lactate and barely adapted to the internal environment of infected macrophages. The V253M in the promoter region has been reported with a high probability of positive selection in MTB (Osório et al., 2013). However, a high mutation rate of Rv1872c was also detected in clinical isolates from PTB, which proved that it is not a unique mutation in CTB isolates. Rv1873 is downstream of Rv1872c, and the Rv1873 product reveals limited similarity to various proteins (Garen et al., 2006). The S3L mutant of Rv1873 was detected only in 3.45% (2/58) of CTB isolates and was not found in any PTB isolates, which suggest the possibility of independent evolution in skin infection. Among the 1676 coding genes analyzed in the burden test, four genes with cumulative missense SNP effects were marked as significantly different in the distribution of the PTB and CTB groups. Rv3829c is a probable phytoene dehydrogenase and may play a key role in the transformation of carotenoids from colorless to colored (Rose et al., 2013). Rv0392c is an NADH dehydrogenase coding gene with a membrane-bound domain and participates in energy metabolism. The product was found with a higher abundance in multidrug-resistant strains and could cause antibiotic resistance if overexpressed (Phong et al., 2015). Rv2088 (*PknJ*) is a transmembrane serine/threonine protein kinase involved in the regulation of pyruvate kinase A and essential for survival under stress conditions (Singh et al., 2014). Rv2331 is a putative nitrate reductase with pH sensitivity activated by the virulence regulator PhoP (Bansal et al., 2017). Notably, all these genes are engaged in mycobacterium metabolism. In particular, Rv3829c

and Rv0392c encode dehydrogenase genes. In this study, strains with mutations in Rv0392c and Rv3829c were found in 36 and 31 samples, respectively. The two mutants overlapped in more than 50% of the strains, and most of the missense variants led to decreased protein stability in prediction (8/10 in Rv0392c, 8/8 in Rv3829c) (Capriotti et al., 2005). Meanwhile, significantly more drug-resistant mutations were found in strains with mutations in the two genes (Rv0392c, 15/36, 41.67%, $\chi^2 = 9.0576$, $p = 0.003$; Rv3829c, 15/31, 41.67%, $\chi^2 = 13.844$, $p < 0.001$). Thus, we suggest that the two genes synergistically exert influence and maybe the fitness cost of antibiotic-resistant mutations and lead to lower adaptability to skin infection.

This study provided the first genomic datasets of isolates from CTB and identified the molecular characteristics of MTB isolated from CTB in southeastern China. The results showed similar genomic composition and genotyping results in strains from the four major types of CTB and PTB, which suggested that the various cutaneous infection types could barely be affected by the genetic differences of MTB. It also suggested the rare but general risk of cutaneous infection in tuberculosis patients. However, we found lower drug mutation rates of CTB isolates and differences in SNPs of metabolism-related genes among the two groups. The lower genetic variation rate in drug-related genes evidenced the low antibiotic resistance in clinical practice and supported 6-month drug therapy (2HRZE + 4HR) recommended by the WHO for CTB. The metabolic genes found by comparative genomic methods and their relationships with drug resistance provide a reference for future studies. The major limitation of this study is a small dataset of CTB isolates compared with PTB cases involved, which may weaken the effect of comparative genome analysis and fail to draw definitive conclusions. This impact could be diminished by including more isolates in future studies. Another limitation is that other potential genomic factors that may contribute to the development of CTB are ignored, which could be improved by enhancing the overall sequence quality of the samples.

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 below: https://www.ncbi.nlm.nih.gov/, PRJNA820632.

Ethics statement

The study involving human participants was reviewed and approved by the ethics committee of Hospital of Skin disease, Institute of Dermatology, Chinese Academy of Medical Sciences and Peking Union Medical College. Written informed consent to participate in this study was obtained from all participants.

Author contributions

HSW contributed to the conception of the study. WYZ, JSX, LW, YQC, SYL, and CP contributed to data collection. YMM, HQJ, and YS performed the experiment. YMM, JYS, and TL contributed to data analysis. YMM wrote the original draft. HSW and TL contributed to review and editing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the National Key Research and Development Program of China (2019YFE0113300), the National Natural Science Foundation of China (Grants 81972950, 82173431, 82103748, and 82203940), and the Nanjing Incubation Program for National Clinical Research Center (2019060001).

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.2023.1165916/full#supplementary-material

SUPPLEMENTARY TABLE 1
Sequencing information of the 58 CTB isolates.

References

Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40, e126. doi: 10.1093/nar/gks406

Allué-Guardia, A., García, J. I., and Torrelles, J. B. (2021). Evolution of drug-resistant *Mycobacterium tuberculosis* strains and their adaptation to the human lung Environment. *Front. Microbiol.* 12, 612675. doi: 10.3389/fmicb.2021.612675

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75. doi: 10.1186/1471-2164-9-75

Bansal, R., Anil Kumar, V., Sevalkar, R. R., Singh, P. R., and Sarkar, D. (2017). *Mycobacterium tuberculosis* virulence-regulator PhoP interacts with alternative sigma factor SigE during acid-stress response. *Mol. Microbiol.* 104, 400–411. doi: 10.1111/mmi.13635

Billig, S., Schneefeld, M., Huber, C., Grassl, G. A., Eisenreich, W., Bange, F. C., et al. (2017). Lactate oxidation facilitates the growth of *Mycobacterium tuberculosis* in human macrophages. *Sci. Rep.* 7, 6484. doi: 10.1038/s41598-017-05916-7

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17, 238. doi: 10.1186/s13059-016-1108-8

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310. doi: 10.1093/nar/gki375

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

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* 6, 80–92. doi: 10.4161/fly.19695

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

Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., et al. (2016). Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate ald in D-cycloserine resistance. *Nat. Genet.* 48, 544–551. doi: 10.1038/ng.3548

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J., et al. (2019). GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat. Commun.* 10, 2128. doi: 10.1038/s41467-019-10110-6

Franco-Paredes, C., Marcos, L. A., Henao-Martínez, A. F., Rodríguez-Morales, A. J., Villamil-Gómez, W. E., Gotuzzo, E., et al. (2018). Cutaneous mycobacterial infections. *Clin. Microbiol. Rev.* 32, e00069–e00018. doi: 10.1128/CMR.00069-18

Garen, C. R., Cherney, M. M., Bergmann, E. M., and James, M. N. (2006). The molecular structure of Rv1873, a conserved hypothetical protein from *Mycobacterium tuberculosis*, at 1.38 A resolution. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* 62(Pt 12), 1201–1205. doi: 10.1107/S1744309106046902

Gomes, T., Reis-Santos, B., Bertolde, A., Johnson, J. L., Riley, L. W., Maciel, E. L., et al. (2014). Epidemiology of extrapulmonary tuberculosis in Brazil: a hierarchical model. *BMC Infect. Dis.* 14, 9. doi: 10.1186/1471-2334-14-9

Hill, M. K., and Sanders, C. V. (2017). Cutaneous tuberculosis. *Microbiol. Spectr.* 5, doi: 10.1128/microbiolspec.TNMI7-0010-2016

Jabir, R. A., Rukmana, A., Saleh, I., and Kurniawati, T. (2017). The *Existence of Mycobacterium tuberculosis in Microenvironment of Bone*. IntechOpen. Available online at: https://www.semanticscholar.org/paper/The-Existence-of-Mycobacterium-tuberculosis-in-of-Jabir-Rukmana/51693ce3b704cbf2f0111762573f954481ab35e0 (accessed July 5, 2023).

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 46, D335–D342. doi: 10.1093/nar/gkx1038

Karin, H. P., Andreas, R. E., Hans-Henrik, S., Torbjørn, R., 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

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

Letunic, I., and Bork, P. (2011). Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39, W475–W478. doi: 10.1093/nar/gkr201

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

Lin, D., Cui, Z., Chongsuvivatwong, V., Palittapongarnpim, P., Chaiprasert, A., Ruangchai, W., et al. (2020). The geno-spatio analysis of *Mycobacterium tuberculosis* complex in hot and cold spots of Guangxi, China. *BMC Infect. Dis.* 20, 462. doi: 10.1186/s12879-020-05189-y

Liu, Q., Yang, D., Xu, W., Wang, J., Bing, L. V., Shao, Y, et al. (2011). Molecular typing of *Mycobacterium tuberculosis* isolates circulating in Jiangsu province, China. *BMC Infect. Dis.* 11, 288. doi: 10.1186/1471-2334-11-288

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1, 18. doi: 10.1186/2047-217X-1-18

Napier, G., Campino, S., Merid, Y., Abebe, M., Woldeamanuel, Y., Aseffa, A., et al. (2020). Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* 12, 114. doi: 10.1186/s13073-020-00817-3

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

Osório, N. S., Rodrigues, F., Gagneux, S., Pedrosa, J., Pinto-Carbó, M., Castro, A. G., et al. (2013). Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. *Mol. Biol. Evol.* 30, 1326–1336. doi: 10.1093/molbev/mst038

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

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. *Microb. Genom.* 2, e000056. doi: 10.1099/mgen.0.000056

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

Phong, T. Q., Ha do, T. T., Volker, U., and Hammer, E. (2015). Using a label free quantitative proteomics approach to identify changes in protein abundance in multidrug-resistant *Mycobacterium tuberculosis*. *Indian J. Microbiol.* 55, 219–230. doi: 10.1007/s12088-015-0511-2

Rose, G., Cortes, T., Comas, I., Coscolla, M., Gagneux, S., Young, D. B., et al. (2013). Mapping of genotype-phenotype diversity among clinical isolates of MTB by sequence-based transcriptional profiling. *Genome Biol. Evol.* 5, 1849–1862. doi: 10.1093/gbe/evt138

Sagulenko, P., Puller, V., and Neher, R. A. (2018). TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* 4, vex042. doi: 10.1093/ve/vex042

Sandgren, A., Hollo, V., and van der Werf, M. J. (2013). Extrapulmonary tuberculosis in the European Union and European Economic Area, 2002 to 2011. *Euro Surveill.* 18, 20431. doi: 10.2807/ese.18.12.20431-en

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689. doi: 10.1093/nar/gki366

Seemann, T. (2023). *Snippy: Fast Bacterial Variant Calling from ngs Reads*. Available online at: https://github.com/tseemann/snippy (accessed July 5, 2023).

Singh, D. K., Singh, P. K., Tiwari, S., Singh, S. K., Kumari, R., Tripathi, D. K., et al. (2014). Phosphorylation of pyruvate kinase A by protein kinase J leads to the altered growth and differential rate of intracellular survival of mycobacteria. *Appl. Microbiol. Biotechnol.* 98, 10065–10076. doi: 10.1007/s00253-014-5859-4

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034. doi: 10.1093/bioinformatics/btv098

World Health Organization (2021). *Global Tuberculosis Report 2021. License: CC BY-NC-SA 3, 0. IGO*. Available online at: https://www.who.int/publications-detail-redirect/9789240037021 (accessed July 5, 2023).

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., et al. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. doi: 10.1093/nar/gkz333

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, 788–795. doi: 10.1016/S1473-3099(18)30218-4

Yang, T., Zhong, J., Zhang, J., Li, C., Yu, X., Xiao, J., et al. (2018). Pan-genomic study of *Mycobacterium tuberculosis* reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Front. Microbiol.* 9, 1886. doi: 10.3389/fmicb.2018.01886

Yao, S., Yan, J., Li, L., Ma, D., Liu, J., Wang, Q., et al. (2020). Determining *Mycobacterium tuberculosis* drug resistance and risk factors for multidrug-resistant tuberculosis in sputum smear-positive tuberculosis outpatients in Anhui Province, China, 2015-2016. *Infect. Drug Resist.* 13, 1023–1032. doi: 10.2147/IDR.S244482

Zhang, J., Fan, Y. K., Wang, P., Chen, Q. Q., Wang, G., Xu, A. E., et al. (2018). Cutaneous tuberculosis in China - a multicentre retrospective study of cases diagnosed between 1957 and 2013. *J. Eur. Acad. Dermatol. Venereol.* 32, 632–638. doi: 10.1111/jdv.14851

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 18 May 2023
DOI 10.3389/fmicb.2023.1175925

Check for updates

OPEN ACCESS

EDITED BY
Paulo Jorge Dias,
University of Lisbon, Portugal

REVIEWED BY
Bowen Song,
University of Liverpool, United Kingdom
Kunqi Chen,
Fujian Medical University, China
Zi Liu,
Nanjing University of Science and Technology, China

*CORRESPONDENCE
Runyu Jing
jingryedu@gmail.com
Jiesi Luo
ljs@swmu.edu.cn

†These authors have contributed equally and share first authorship

RECEIVED 28 February 2023
ACCEPTED 27 April 2023
PUBLISHED 18 May 2023

CITATION
Yu L, Zhang Y, Xue L, Liu F, Jing R and Luo J (2023) Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy. *Front. Microbiol.* 14:1175925. doi: 10.3389/fmicb.2023.1175925

COPYRIGHT
© 2023 Yu, Zhang, Xue, Liu, Jing and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy

Lezheng Yu[1†], Yonglin Zhang[2†], Li Xue[3], Fengjuan Liu[4], Runyu Jing[5*] and Jiesi Luo[6,7*]

[1]School of Chemistry and Materials Science, Guizhou Education University, Guiyang, China, [2]Department of Pharmacy, Affiliated Hospital of North Sichuan Medical College, Nanchong, China, [3]School of Public Health, Southwest Medical University, Luzhou, China, [4]School of Geography and Resources, Guizhou Education University, Guiyang, China, [5]School of Cyber Science and Engineering, Sichuan University, Chengdu, China, [6]Basic Medical College, Southwest Medical University, Luzhou, China, [7]Sichuan Key Medical Laboratory of New Drug Discovery and Druggability Evaluation, Luzhou Key Laboratory of Activity Screening and Druggability Evaluation for Chinese Materia Medica, Southwest Medical University, Luzhou, China

Post-transcriptionally RNA modifications, also known as the epitranscriptome, play crucial roles in the regulation of gene expression during development. Recently, deep learning (DL) has been employed for RNA modification site prediction and has shown promising results. However, due to the lack of relevant studies, it is unclear which DL architecture is best suited for some pyrimidine modifications, such as 5-methyluridine (m^5U). To fill this knowledge gap, we first performed a comparative evaluation of various commonly used DL models for epigenetic studies with the help of autoBioSeqpy. We identified optimal architectural variations for m^5U site classification, optimizing the layer depth and neuron width. Second, we used this knowledge to develop Deepm5U, an improved convolutional-recurrent neural network that accurately predicts m^5U sites from RNA sequences. We successfully applied Deepm5U to transcriptomewide m^5U profiling data across different sequencing technologies and cell types. Third, we showed that the techniques for interpreting deep neural networks, including LayerUMAP and DeepSHAP, can provide important insights into the internal operation and behavior of models. Overall, we offered practical guidance for the development, benchmark, and analysis of deep learning models when designing new algorithms for RNA modifications.

KEYWORDS

RNA 5-methyluridine, deep learning, autoBioSeqpy, UMAP, SHAP

Introduction

To date, over 170 types of chemical modifications have been identified in cellular RNAs, which contain not only some common types such as N^6-methyladenosine (m^6A), 5-methylcytosine (m^5C), N^1-methyladenosine (m^1A), pseudouridine (Ψ), 5-hydroxymethylcytosine (5hmC), and 2'-O-methylation of ribose (2'-O-Me), but also several rare types, including 7-methylguanosine (m^7G), adenosine-to-inosine (A-to-I), dihydrouridine (D), N^2-methylguanosine (m^2G), and N4-acetylcytidine (ac4C; El Allali et al., 2021). All four RNA bases, as well as the ribose sugar, can be the targets for modification, and

almost all RNA species are modified, with transfer RNA (tRNA) and ribosomal RNA (rRNA) being the most heavily modified (Roundtree et al., 2017; Chen et al., 2020). RNA modifications affect numerous biological processes, including regulation of post-transcriptional gene expression, mRNA life cycle, RNA localization, ncRNA biogenesis and function (Alarcón et al., 2015; Meyer et al., 2015; Nachtergaele and He, 2018). Accordingly, aberrant modifications are widely lined to developmental disease (Jonkhout et al., 2017). Increasing evidence suggests that RNA modification pathways are also misregulated in cancers and may be ideal targets for cancer therapy (Delaunay and Frye, 2019; Barbieri and Kouzarides, 2020).

Recent advances in studying RNA modifications have been benefited tremendously from improved detection methods. Liquid chromatography coupled with mass spectrometry (LC–MS) and next-generation sequencing (NGS) are two main methodologies for identifying and quantifying RNA modifications (Wiener and Schwartz, 2021). The LC–MS allows direct measurement of many modifications with excellent sensitivity and specificity, but is limited in its ability to determine sequence context (Su et al., 2014; Wetzel and Limbach, 2016). In contrast to LC–MS, high-throughput sequencing can provide information about the sequence context of long RNAs, which facilitates the detection of modifications in a transcriptome-wide manner (Li et al., 2016). However, most RNA modifications cannot be directly detected by NGS-based approaches, because all RNA-sequencing NGS library generation protocols include a reverse transcription step where RNA is converted into DNA (Sarkar et al., 2021). This step is sensitive to specific RNA modifications that can slow down or block the reverse transcriptase or induce the misbinding of nucleotides in the cDNA (Suzuki et al., 2015).

Owing to the high cost and technical challenge of experimentally detecting all possible modification candidates, researchers have attempted to computationally identify the epitranscriptome. Most modern computational approaches use machine learning (ML) algorithms based on handcrafted features to train a predictive model (Zhou et al., 2016; Chen et al., 2019). Although such models seem to be more transparent and controllable in construction, the bias of user assumptions in feature engineering limits their performance. In keeping with the general trend in artificial intelligence (AI), there has been a switch from classical machine learning to deep learning in newly developed RNA modification predictors. For instance, the m^6A site predictors DeepM6ASeq (Zhang and Hamada, 2018), PM6ACNN (Alam et al., 2020), and DNN-m6A (Zhang et al., 2021b); Ψ site predictors iPseU-CNN (Tahir et al., 2019a), MU-PseUDeep (Khan et al., 2020), and PseUdeep (Zhuang et al., 2021); 5hmC site predictor iRhm5CNN (Ali et al., 2021); 2'-O-Me site predictors Deep-2'-O-Me (Mostavi et al., 2018), iRNA-PseKNC (2methyl) (Tahir et al., 2019b), and DeepOMe (Li et al., 2021); ac4C site predictors DeepAc4C and CNNLSTMac4CPred (Wang et al., 2021; Zhang et al., 2022); and a disease-associated m^7G site predictor HN-CNN (Zhang et al., 2021a). A strength of these predictors is that they can learn modification determinants directly from sequencing data, avoiding biased user-defined features. Thus, many DL methods outperform classical ML approaches in benchmarks with different RNA modifications (Tahir et al., 2019a; Wang et al., 2021). Despite its successes, deep learning also poses challenges and limitations. First, the accessibility remains riddled with technical challenges for the nonexpert users. As most DL methods are available as source code, running them proficiently requires advanced knowledge specific to the field. Second, due to the complexity of network architectures and large training parameters, DL models are often treated as black boxes that simply mapping a given input to a model output without the explanation of how and why they work.

As another critical and abundant epigenetic mark, the 5-methyluridine (m^5U) modification has attracted the attention of researchers worldwide. This modification is not only frequently detected in cytosolic tRNAs (Carter et al., 2019; Powell and Minczuk, 2020), but also found in other non-coding RNAs such as mRNA and rRNA (Phizicky and Alfonzo, 2010; Keffer-Wilkes et al., 2020). Some typical enzymes are involved in the catalytic procedure of m^5U modification in different organisms, including RlmC, RlmD, and TrmA in *Escherichia coli* (Urbonavicius et al., 2007; Powell and Minczuk, 2020), Trm2 in *Saccharomyces cerevisiae* (Nordlund et al., 2000), and TRMT2A and TRMT2B (Sequence homology to TrmA and Trm2 respectively) in mammals (Carter et al., 2019; Jiang et al., 2020; Pereira et al., 2021). For this modification, the conserved T-loop motif has been found in various RNAs, which plays an important role in stabilizing the tertiary structure of RNAs (Powell and Minczuk, 2020; Pereira et al., 2021). To clarify its biological functions and understand the relevant biological processes, there is an urgent need to accurately identify RNA m^5U sites.

Some experimental and computational methods have been developed for this mark, such as FICC-Seq, miCLIP-Seq, m5UPred and RNA-m5U (Carter et al., 2019; Jiang et al., 2020; Feng and Chen, 2022). More recently, the RNA domain separation network (RNADSN) has been proposed to abstract common features between tRNA and mRNA m5U modifications to improve the prediction of m5U sites, which mixes several layers from the convolutional neural network (CNN) and long short-term memory (LSTM; Li et al., 2022). However, studies on the identification and functional characterization of m^5U remain limited and unexplored in current literature, and a further study on the application of deep learning in m^5U prediction is still very necessary and useful.

In the present study, we explore the use of state-of-art DL algorithms and advanced interpretable techniques, and propose a novel computational tool for rapidly and accurately identifying RNA m5U sites. In ordor to save calculation time and make direct comparisons, only the one-hot encoding method was utilized to code RNA sequences here. Five different DL architectures such as the convolutional neural network (CNN), recurrent neural networks (RNNs) with bidirectional long short-term memory (BiLSTM) or bidirectional gated recurrent units (BiGRU), and the combination of the two networks (CNN-BiLSTM and CNN-BiGRU), were employed to build the DL models. Experimental results showed that the CNN-BiLSTM model achieved the best overall prediction performance on both of the *Full_train* and *Full_test* datasets, providing the highest scores of *ACC* (92.32 and 92.91%), and *MCC* (0.8465 and 0.8584). When performing on the cross-cell-type and cross-technique validation, the CNN-BiLSTM model also obtained satisfactory prediction results, and was eventually named Deepm5U. Using the same datasets, the predictive performance of Deepm5U was superior to that of the exiting method. Furthermore, our Deepm5U was visualized to understand how the model is processing information and making decisions. Finally, we used autoBioSeqpy (Jing et al., 2020) to

develop, train, and assess different DL models, and offered a step-by-step guide on how to execute them.

Materials and methods

Benchmark datasets

A high-quality dataset is essential for developing a reliable prediction model. Currently, there are several public databases focused on RNA modifications, including versatile database for multiple modification types, such as RMBase (Sun et al., 2016; Xuan et al., 2018), MODOMICS (Boccaletto et al., 2018, 2022), and RMVar (Luo et al., 2021), and specialized database for a particular modification type, like Met-DB (Liu et al., 2015, 2018, 2021), REPIC (Liu et al., 2020), m^{6}A-Atlas (Tang et al., 2021), CVm6A (Han et al., 2019), m6AVar (Zheng et al., 2018), m5C-Atlas (Ma et al., 2022) and m7GHub (Song et al., 2020). Unfortunately, until now, there has not been such a database available for m^{5}U data. Therefore, we chose a published benchmark dataset constructed by Jing et al. (2020) to develop our deep learning models. Experimentally validated m^{5}U sites (positive samples) were generated by integrating the sequencing results of FICC-seq and miCLIP-seq technologies on two cell lines, HEK293 and HAP1. The same number of unmodified uridine sites (non-m^{5}U, negative samples) were randomly sampled from the same transcripts of positive samples. All samples were 41 nt in length, with the modified and unmodified uridine sites located in the center of these sequences. Based on the genomic location, positive and negative m^{5}U sites were further divided into two categories: full transcript mode (uridine sites located in both exonic and intronic regions) and mature mRNA mode (uridine sites only located on mature mRNA transcripts). For each mode, total sequences were split into two mutually exclusive datasets: a training dataset of ~80% of the instances used for model derivation and an independent test set of the remaining 20% used to evaluate model accuracy (Supplementary Table S1).

In addition, Jing et al. (2020) separated the benchmark dataset into eight subsets, namely HAP1_full, HEK293_full, FICC_full, miCLIP_full, HAP1_mature, HEK293_mature, FICC_mature, and miCLIP_mature, to investigate the effects of two experimental parameters, sequencing technique, and cell type, on model prediction performance. Leave-one-subset-out cross-technique and cross-cell-type validations were performed by repeating the training-test procedure iteratively such that each subset was used as the test set exactly once. For instance, when training with the HAP1_full or HAP1_mature, the performance of model was evaluated by the HEK293_full or HEK293_mature, and vice versa. Notably, all subsets were constructed with a 1:10 positive-to-negative ratio (Supplementary Table S2).

Deep learning techniques

Deep learning, so far the most successful form of machine learning, uses a synthetic neural network architecture composed of multiple sequential layers that can be trained on input data to achieve a prediction task. The idea of deep learning is that stacks of simple layers can learn end to end, automatically discovering a higher-level representation of the original data, which is extremely powerful and flexible in a variety of relationships that they can model (Wainberg et al., 2018). Various types of network layers, such as convolution layers, pooling layers, recurrent layers, activation layers, and fully connected layers, have been proposed to support the construction of highly flexible model architectures. CNN layers use convolution operations to fuse features that are close to each other and transfer them by kernels:

$$Convolution\left(X\right)_{ik} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{mn}^{k} x_{i+m,n} \tag{1}$$

where X is the layer input, i and k are the indices of output position and filter kernel, respectively (LeCun et al., 2015). Convolutional filter w^k is the $M \times N$ weight matrix with M and N being the window size and input channel, respectively. Additionally, the convolution operation can be adapted to a wider range of information fusion by changing the padding size and dilation size. For sequential data, CNN can fusion the environment of each base so that the bases differ depending on their neighbors. Sometimes similar performance can be achieved using K-mer or sliding window operations, but using CNNs can result in lower sparsity and an editable number of channels for further processing of the data. The pooling layer usually follows the convolution layer, and its function is to downsample layer's input by computing the maximum or average value of the features over a region. In the RNN family, the layers use each unit of the sequence to update the hidden state for learning and inference from context. In the recurrent layer, tensors are used to represent the hidden state, and each unit of the sequence will be encoded by one or more fully connected layers for updating the hidden cells. For example, the long short-term memory (LSTM) layer contains hidden states and cell states that are updated in each iteration:

$$f_t = \sigma\left(W_{fh} h_{t-1} + W_{fx} x_t + b_f\right) \tag{2}$$

$$i_t = \sigma\left(W_{ih} h_{t-1} + W_{ix} x_t + b_i\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(W_{Ch} h_{t-1} + W_{Cx} x_t + b_c\right) \tag{4}$$

$$C_t = f_t C_{t-1} + i_t \times \tilde{C}_t \tag{5}$$

$$o_t = \sigma\left(W_{oh} h_{t-1} + W_{ox} x_t + b_o\right) \tag{6}$$

$$h_t = o_t \times \tanh\left(C_t\right) \tag{7}$$

Where C_t and h_t are the cell state and hidden state, respectively (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Jin et al., 2021). Whereas the gate recurrent unit (GRU) layer only updates the hidden state:

$$r_t = \sigma\left(W_{rh}h_{t-1} + W_{rx}x_t + b_r\right) \tag{8}$$

$$\tilde{h}_t = \varnothing_h\left(W_{hh}\left(r_t \times h_{t-1}\right) + W_{hx}x_t + b_h\right) \tag{9}$$

$$z_t = \sigma\left(W_{zh}h_{t-1} + W_{zx}x_t + b_z\right) \tag{10}$$

$$h_t = \left(1 - z_t\right) \times \tilde{h}_t + z_t \times h_{t-1} \tag{11}$$

In addition, bidirectional operation, which allows RNN layers to learn a sequence starting from both head and tail directions, since RNNs process sequences without a predetermined direction. Dense layer, also known as fully connected layer, is the simplest type of layer, where every input is connected to every output. The role of activation function is to introduce the nonlinearity in the input–output relationship. Frequently used activation functions include sigmoid and rectified linear unit (ReLU), which are given by:

$$\text{Sigmoid}\left(x\right) = \frac{1}{1 + e^{-x}} \tag{12}$$

$$\text{ReLu}\left(x\right) = \begin{cases} x \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases} \tag{13}$$

Usually, ReLU is used for the nonlinear gain of the model, while sigmoid is used in the output layer of the binary classification model.

Model architectures

We designed five architectures, namely, CNN, BiLSTM, BiGRU, CNN-BiLSTM and CNN-BiGRU, which use 20 nucleotides on each side of a position of interest as input, and output the probability of the position being an m^{5}U site and a non-m^{5}U site. The input to the models is a sequence of one-hot encoded nucleotides, where A, C, G, and U are encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1] respectively and the output of the models is a score in the range [0, 1], representing positive (T or 1) and negative (F or 0) classes. We run a grid search to exhaustively test the combinations of convolution layers (1, 2, 3), kernel size (3, 5, 7, 9, 11), number of filters (50, 150, 250), pool size (2, 4, 6, 8, 10), LSTM layers (1, 2, 3), number of units in the LSTM layer (32, 64, 128, 256), GRU layers (1, 2, 3), and number of units in the GRU layer (32, 64, 128, 256) to select the best hyperparameters for the models. Details about optimal hyperparameters and model architectures are provided in Supplementary Figure S1; Supplementary Tables S3–S5 and below, respectively.

Convolutional neural network architecture (CNN).

(1) Convolution layer (250 filters; kernel size, 11; ReLU activation; 0% dropout; step size, 1)
(2) Convolution layer (250 filters; kernel size, 11; ReLU activation; 0% dropout; step size, 1)
(3) Pooling layer (maximum value; pool size, 10, step size, 10)
(4) Fully connected layer (256 units)
(5) Dropout layer (20% dropout)
(6) Activation layer (ReLU activation)
(7) Output layer (1 units, sigmoid activation)

Bidirectional long short-term memory architecture (BiLSTM).

(1) Bidirectional LSTM layer (256 units, tanh activation; sigmoid recurrent activation; 0% dropout)
(2) Bidirectional LSTM layer (256 units, tanh activation; sigmoid recurrent activation; 0% dropout)
(3) Fully connected layer (256 units)
(4) Dropout layer (20% dropout)
(5) Activation layer (ReLU activation)
(6) Output layer (1 unit, sigmoid activation)

Bidirectional gated recurrent unit architecture (BiGRU).

(1) Bidirectional GRU layer (256 units, tanh activation; sigmoid recurrent activation; 0% dropout)
(2) Fully connected layer (256 units)
(3) Dropout layer (20% dropout)
(4) Activation layer (ReLU activation)
(5) Output layer (1 unit, sigmoid activation)

Convolutional-bidirectional long short-term memory architecture (CNN-BiLSTM).

(1) Convolution layer (250 filters; kernel size, 7; ReLU activation; 0% dropout; step size, 1)
(2) Convolution layer (250 filters; kernel size, 7; ReLU activation; 0% dropout; step size, 1)
(3) Pooling layer (maximum value; pool size, 4, step size, 4)
(4) Bidirectional LSTM layer (64 units, tanh activation; sigmoid recurrent activation; 0% dropout)
(5) Fully connected layer (256 units)
(6) Dropout layer (20% dropout)
(7) Activation layer (ReLU activation)
(8) Output layer (1 unit, sigmoid activation)

Convolutional-bidirectional gated recurrent unit architecture (CNN-BiGRU).

(1) Convolution layer (250 filters; kernel size, 11; ReLU activation; 0% dropout; step size, 1)
(2) Pooling layer (maximum value; pool size, 10, step size, 10)
(3) Bidirectional GRU layer (256 units, tanh activation; sigmoid recurrent activation; 0% dropout)
(4) Fully connected layer (256 units)

(5) Dropout layer (20% dropout)
(6) Activation layer (ReLU activation)
(7) Output layer (1 unit, sigmoid activation)

Model training

All models were trained using the Adam optimizer with a learning rate of 0.001, epoch of 20 and batch size of 64. During training, the sequences were first re-shuffled and subsequently randomly split into training (70%), validation (10%) and testing (20%) fractions. The validation set was used to evaluate the binary cross-entropy loss after each epoch, and the test set was used to evaluate the model. For each architecture, we repeated the training procedure 5 times and used the average result of five trained models as the final prediction. To implement the models, we used the autoBioSeqpy software with Keras framework (Chollet, 2015) and trained them on a standard PC equipped an Intel Core i7-9700K CPU, 16GB working memory and a 16 GB NVIDIA GeForce RTX 2070 GPU.

Model interpretation and visualization

We tried to interpret the DL models by visualizing the manifold of intermediate outputs and measuring the contribution of the inputs. Currently, uniform manifold approximation and projection (UMAP) library is available for projecting a high-dimensional layer into lower dimension (usually 2D for visualization) while keeping the distances of every pair of samples as possible (McInnes and Healy, 2018). To better visualize the outputs of hidden layers, we integrated the UMAP library into LayerUMAP, a new plugin for autoBioSeqpy. Using LayerUMAP, we can generate the manifold projection of any hidden layer and observe the evolution of internal representation layer by layer during the training process. Similarly, we integrated SHAP (SHapley Additive exPlanations) into autoBioSeqpy to develop DeepSHAP for measuring the contribution of sequence inputs. SHAP is an implementation of computing shapely values, which is a solution concept in game theory:

$$\begin{aligned}\phi_i(f,x) &= \sum_{S\subseteq S_{all}/\{i\}} \frac{|S|!(M-|S|-1)!}{M!}\left[f_x(S\cup\{i\})-f_x(S)\right] \\ &= \sum_{S\subseteq S_{all}/\{i\}} \frac{1}{C_M^{|S|}(M-|S|)}\left[f_x(S\cup\{i\})-f_x(S)\right] \end{aligned} \quad (14)$$

where M is the number of features, S is a subset of the features, f is the model, $S_{all}/\{i\}$ is all the possible subset exclude feature i, and f_x is the conditional expectation function. The total contribution of features can be represented by a linear combination of Shapley value:

$$f(x)=\phi_0(f)+\sum_{i=1}^{M}\phi_i(f,x) \quad (15)$$

where $\phi_0(f)=f_x(\varnothing)$. DeepSHAP can visualize the SHAP values of input sequences using the heat maps or logo plots for the downstream analysis.

Evaluation metrics

For evaluation, we calculated the accuracy (*ACC*), precision (*PRE*), *F*-value, recall, and Matthew's correlation coefficient (*MCC*) as quantification metrics, which are defined as follows:

$$ACC=\frac{TP+TN}{TP+FP+TN+FN} \quad (16)$$

$$PRE=\frac{TP}{TP+FP} \quad (17)$$

$$F-value=2\times\frac{TP}{2TP+FP+FN} \quad (18)$$

$$Recall=\frac{TP}{TP+FN} \quad (19)$$

$$MCC=\frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \quad (20)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive and false negative, respectively. Moreover, we plotted receiver operating characteristic (ROC) and precision-recall (PR) curves, and summarized model performance by computing areas under both ROC and PR curves, resulting in auROC and auPR, respectively.

Overview of autoBioSeqpy

The autoBioSeqpy is a Keras-based deep learning software for fast and easy development, training, and analysis of deep learning model architectures for biological sequence classification (Jing et al., 2020). Compared with other tools or libraries, the biggest advantage of autoBioSeqpy is its simplicity and flexibility, which is especially suitable for nonexperts or users with little or no knowledge of deep learning techniques. No programming required, users only need to prepare input datasets and model templates. The operation is also simple. The entire workflows, including file reading, data encoding, parameter initialization, model training, evaluation, and visualization, can be automatically executed with just one-line instruction.

Results

Evaluation of representative DL models

Five representative DL models constructed with different network architectures were used for benchmarking ("Model architectures" section). Using the instruction 1 as shown in Figure 1, we first assessed basic prediction performance of each model on full transcript mode

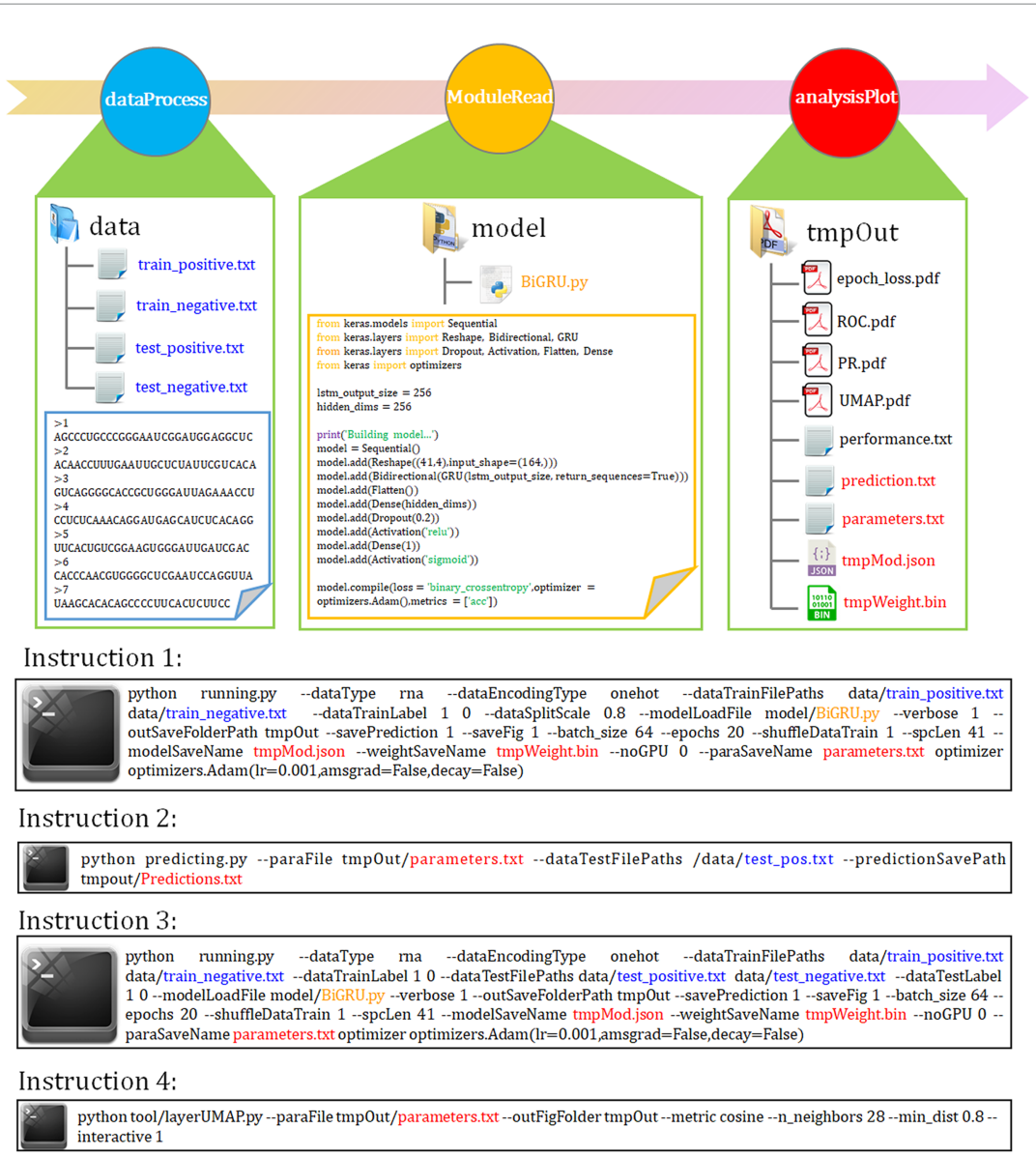


FIGURE 1
autoBioSeqpy workflow and usage. The basic framework of autoBioSeqpy consists of six modules, including four self-write modules (*ParaParser*, *dataProcess*, *moduleRead* and *analysisPlot*) and two dependent modules (*Keras* and *sklearn*). The *dataProcess* module encodes DNA, RNA, protein and compound character sequences into model-readable numerical vectors. The *moduleRead* module loads and initializes the neural network model designed by the users. The *analysisPlot* module evaluates the models on the test set. The following are some commonly used commands of autoBioSeqpy, including model training, prediction and visualization.

using a stratified random sampling strategy. With the parameter "--dataSplitScale" set to 0.8, autoBioSeqpy randomly split the input sequences into 80% training-validation set and 20% test set, stratified by class. We used the shuffle mechanism (−-shuffleDataTrain 1) for each call of instruction 1 to avoid overfitting and to ensure that the model demonstrated differences. For each model, we repeated the instruction 1 five times to estimate mean and standard deviation of the seven metrics ("Evaluation metrics" section). Full results were listed and shown in Table 1; Supplementary Figure S2A, respectively. Overall, all models performed well in the intra-dataset evaluation. CNN model showed slightly worse performance compared with hybrid models, but performed better than individual RNN models. The hybrid CNN-BiLSTM model achieved the best prediction performance and provided the highest scores for *ACC* (92.32%), *F*-value (92.29%), *MCC* (0.8465), *auROC* (0.9740), and *auPR* (0.9781). CNN-BiGRU afforded the highest *Recall* score (92.46%), while CNN offered the highest *PRE* score (93.04%). In addition, an independent test set was employed to evaluate the robustness and reproducibility of the presented DL models. Here, both instruction 2 and instruction 3 can be used for this purpose, but they are suitable for different application conditions (Figure 1). For example, we can combine instruction 1 and instruction 2 to predict the probability and class of unlabelled datasets or single-labelled datasets, while for the datasets containing both positive and negative labels, we can directly call instruction 3 to generate their assessment results together with the related plots and confusion matrix. In this comparison, we see that CNN-BiLSTM still achieved the best

TABLE 1 Performance comparison of different deep learning models on the full transcript mode.

Model	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)	MCC	auROC	auPR
Training dataset (Full_train)							
CNN	91.56	91.50	90.04	**93.04**	0.8319	0.9637	0.9690
BiLSTM	88.59	88.55	87.58	89.65	0.7730	0.9419	0.9477
BiGRU	87.22	87.27	87.04	87.60	0.7820	0.9536	0.9544
CNN-BiLSTM	**92.32**	**92.29**	91.93	92.65	**0.8465**	**0.9740**	**0.9781**
CNN-BiGRU	91.85	91.94	**92.46**	91.44	0.8370	0.9632	0.9676
Independent test set (Full_test)							
CNN	91.67	91.63	91.25	92.05	0.8336	0.9689	0.9738
BiLSTM	88.39	88.53	88.60	88.47	0.7676	0.9560	0.9584
BiGRU	88.63	89.01	92.06	86.22	0.7751	0.9631	0.9646
CNN-BiLSTM	**92.91**	**92.96**	93.79	**92.16**	**0.8584**	**0.9773**	**0.9810**
CNN-BiGRU	91.87	92.07	**94.39**	89.95	0.8393	0.9749	0.9788

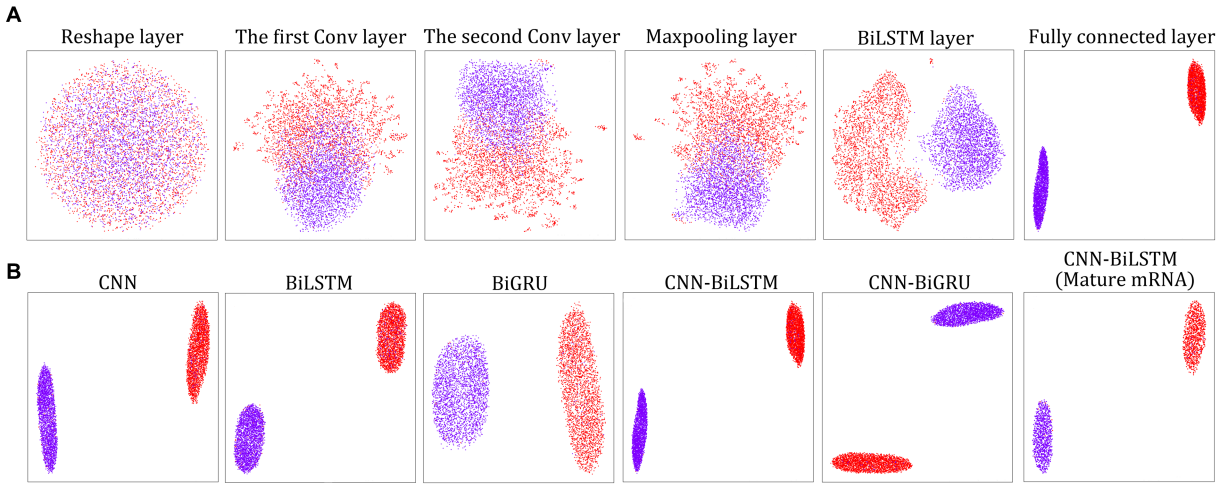


FIGURE 2
LayerUMAP reveals the inner working mechanism of the DL models. **(A)** UMAP projection of layer-to-layer evolution of the CNN-BiLSTM model for the full transcript training dataset. Colored point clouds represent m^{5}U and non-m^{5}U sites, showing how the model clusters the categories. **(B)** Comparison of the UMAP maps of the last hidden layer representation for other DL models and datasets.

performance, followed by CNN-BiGRU, CNN, BiGRU and BiLSTM, which was consistent with the results of the training dataset (Table 1; Supplementary Figure S2B). Next, we benchmarked the performance of above five models using the mature mRNA mode. Like the full transcript mode, no models achieved equivalent performance to CNN-BiLSTM. On average, its prediction accuracy was 91.12% for training dataset and 92.07% for independent test set (Supplementary Table S6). The ROC and PR curves were illustrated in Supplementary Figure S3.

To further assess the predictive performance of CNN-BiLSTM, 41 m^{5}U sites identified by Oxford Nanopore Techniques (ONT) have been collected from the DirectRMDB database (Zhang et al., 2023) as the second independent test set. As shown in supplementary Table S7, 36 of them are correctly predicted by CNN-BiLSTM, and only 5 are misclassified. Hence, CNN-BiLSTM achieves a satisfactory *ACC* (87.80%) once again. Taken together, CNN-BiLSTM constantly outperformed the other DL models and therefore was chosen as the core classification algorithm for Deepm5U, a new predictor for the m^{5}U identification.

Full view of interpretable DL models from hidden layers and input features

Immediately following instruction 1 or instruction 3, we can use the layerUMAP tool to visually investigate the trained DL models via instruction 4 (Figure 1). By default, layerUMAP outputs the last hidden layer projection of the model. However, using the "--interactive" parameter, we can project any hidden layer by specifying the name or index of the layer using a list provided by layerUMAP. We first dissected the best CNN-BiLSTM model layer by layer to verify its ability to distinguish between m^{5}U and non-m^{5}U sites (Figure 2A).

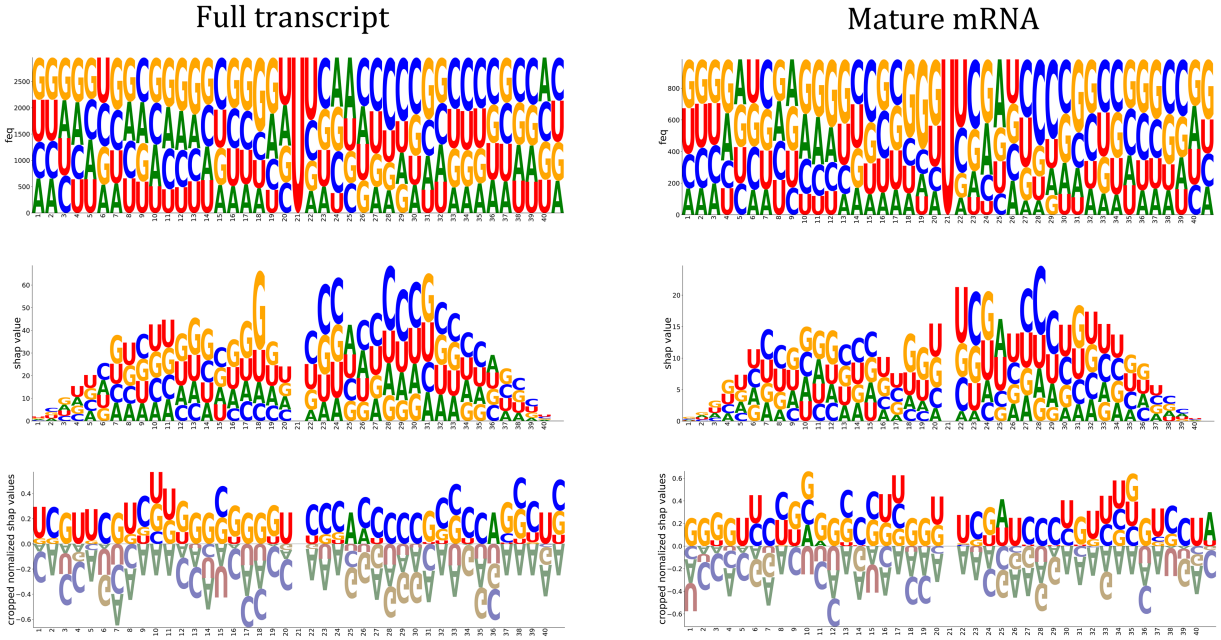


FIGURE 3
Feature importance analysis. Sequence logos representing the nucleotide composition at each position of the input sequence (Top panel). Feature importance scores associate with m^{5}U identification determined by DeepSHAP (CNN-BiLSTM classifier). The height of the letter indicates the SHAP value of the relevant feature for the input sequence (Middle panel). Positive and negative normalized SHAP values represent favored and disfavored relevant features as shown in the Bottom panel.

We analyzed features extracted from six hidden layers of the classification model. First, the features in the reshape layer (original one-hot encoding) were completely mixed and indistinguishable. Along the hierarchy of layers, the features became clearly distinguishable, and in the last two layers (BiLSTM layer and fully connected layer), features separated point populations into two distinct clusters according to their labels. These results demonstrated the powerful feature extraction capability of deep learning algorithm that can automatically extract the useful information from raw inputs. It is worth noting that some hidden layers, such as reshape layer, convolution layer and pooling layer, cannot be projected directly into 2D space for visualization due to their multidimensional data. Therefore, we performed dimensionality reduction on these layers to compress the multi-dimensional data into low-dimensional data from different directions. Figure 2B showed the projection of the last hidden layer for other models and datasets. Unsurprisingly, UMAP visually revealed two-point clusters that correspond to m^{5}U and non-m^{5}U sites, which were in line with the performance of the corresponding DL models.

One of the Jupyter notebook tutorials in the autoBioSeqpy demonstrated how to use DeepSHAP to measure and visualize the contribution of input sequences to a trained DL model.[1] We used the logo plots generated by the Logomaker package instead of the commonly used summary violin plots to display the computed SHAP values (Kim et al., 2020; Tareen and Kinney, 2020). These logos allowed visualization of how important a certain nucleotide at a certain position was for the model's classification decision. We first generated the classical sequence logos for the full-length input sequences to reveal the potential *cis*-regulatory patterns of m^{5}U (Figure 3, top panel). Using 2,956 and 985 training sequences, we calculated nucleotide compositions for the full transcript mode and mature mRNA mode, respectively. We did not observe a significant difference in the motifs between the two modes. Guanine was overrepresented in the upstream region relative to the m^{5}U sites, while some positions in the downstream region were enriched for cytosine. The feature importance scores associated with m^{5}U identification determined by DeepSHAP were shown in the middle panel of Figure 3. Inspired by sequence logos, the height of each letter corresponded to the SHAP value of that base. Since uracil was located at the center position of all samples, its contribution to the overall prediction was zero. Nucleotides near the center contributed more to the prediction (high SHAP values), while nucleotides located on both sides had low SHAP values. GGU at positions {18–20} and CXAXCCC at positions {23–29} made a significant contribution to predicting m^{5}U for both modes. This observation also coincided with the above sequence analysis. Finally, we normalized the SHAP values to highlight the favored and disfavored features. The calculated SHAP values were first scaled to the range of [−0.25, 0.25] to confirm that the summation values could lie in the range of [−1, 1], and then these values were accumulated according to the position of the base and plotted on the bottom panel of Figure 3. It was clear that adenine was a disfavored sequence feature for m5U recognition, regardless of the mode.

1 https://github.com/jingry/autoBioSeqpy/blob/2.0/notebook/Understanding%20the%20contributions%20from%20the%20inputs%20using%20shaps%20(onehot%20case).ipynb

TABLE 2 Cross-cell-type and cross-technique validation using Deepm5U.

Mode	Testing method	Evaluation metric	Cross-technique validation		Cross-cell-type validation	
			miCLIP-Seq	FICC-Seq	HEK293	HAP1
Full transcript	Intra-dataset evaluation	*ACC* (%)	98.31	97.41	98.21	97.61
		F-value (%)	90.16	84.84	89.77	85.86
		Recall (%)	87.09	79.90	87.01	81.50
		PRE (%)	93.49	90.45	92.76	90.97
		MCC	0.8932	0.8362	0.8886	0.8478
		auROC	0.9769	0.9729	0.9858	0.9770
		auPR	0.9387	0.8975	0.9460	0.9269
	Inter-dataset evaluation	*ACC* (%)	95.64	93.90	97.10	94.33
		F-value (%)	71.79	54.02	82.31	58.60
		Recall (%)	61.05	39.44	74.29	44.13
		PRE (%)	87.57	85.93	92.40	87.22
		MCC	0.7093	0.5574	0.8135	0.5967
		auROC	0.9324	0.8332	0.9727	0.8700
		auPR	0.7958	0.6079	0.9062	0.6627
Mature mRNA	Intra-dataset evaluation	*ACC* (%)	99.36	98.30	99.38	98.22
		F-value (%)	96.54	89.94	96.51	90.15
		Recall (%)	94.62	87.80	94.89	89.04
		PRE (%)	98.56	92.47	98.22	91.41
		MCC	0.9621	0.8913	0.9619	0.8922
		auROC	0.9999	0.9820	0.9928	0.9856
		auPR	0.9998	0.9466	0.9863	0.9641
	Inter-dataset evaluation	*ACC* (%)	98.28	95.06	98.62	94.88
		F-value (%)	90.37	64.02	92.52	61.03
		Recall (%)	89.14	48.41	93.90	44.77
		PRE (%)	91.68	94.81	91.32	97.81
		MCC	0.8945	0.6574	0.9181	0.6408
		auROC	0.9931	0.8893	0.9943	0.8733
		auPR	0.9056	0.7014	0.9508	0.6951

Performance evaluation of Deepm5U on the cross-techniques and cross-cell-type datasets

While it is important to evaluate classification performance within a dataset (intra-dataset), realistic scenarios where classifiers are useful require to be evaluated across datasets (inter-dataset). We used eight datasets ("Benchmark datasets" section) to test the Deepm5U' ability to predict m^5U sites. We evaluated the classification performance when training on one dataset and testing on the other. Within intra-dataset predictions, we observed very high prediction accuracy, with *ACC* larger than 97.00% for all datasets (Table 2). However, high accuracy does not guarantee good predictive performance of model, especially with the presence of extremely unbalanced sample ratios in these datasets. We therefore examined other metrics that are more sensitive to sample imbalance, such as Recall. The average Recall of Deepm5U prediction was 87.73%, and 6 of 8 datasets were predicted with recall of at least 85.00%. The lowest Recall score was 79.90%, obtained by FICC_full dataset. These results confirmed good predictive accuracy of Deepm5U in identifying m^5U sites. Supplementary Figure S4 showed the ROC and PR curves evaluated per sequencing protocol and per cell type for two modes. Deepm5U achieved an average auROC of 0.9841 and an auPR of 0.9507 on different classification tasks. We also visualized the positive and negative samples for all datasets using LayerUAMP based on the features learned in the last hidden layer. As shown in Supplementary Figure S5, Deepm5U mapped input sequences into different clusters according to the two-class label, and we see that the structures of red and purple classes were similar for all cases.

SHAP values explained the bias observed in cross-evaluation

When evaluating the performance of Deepm5U across datasets, we found that the performance of different datasets varied greatly, and

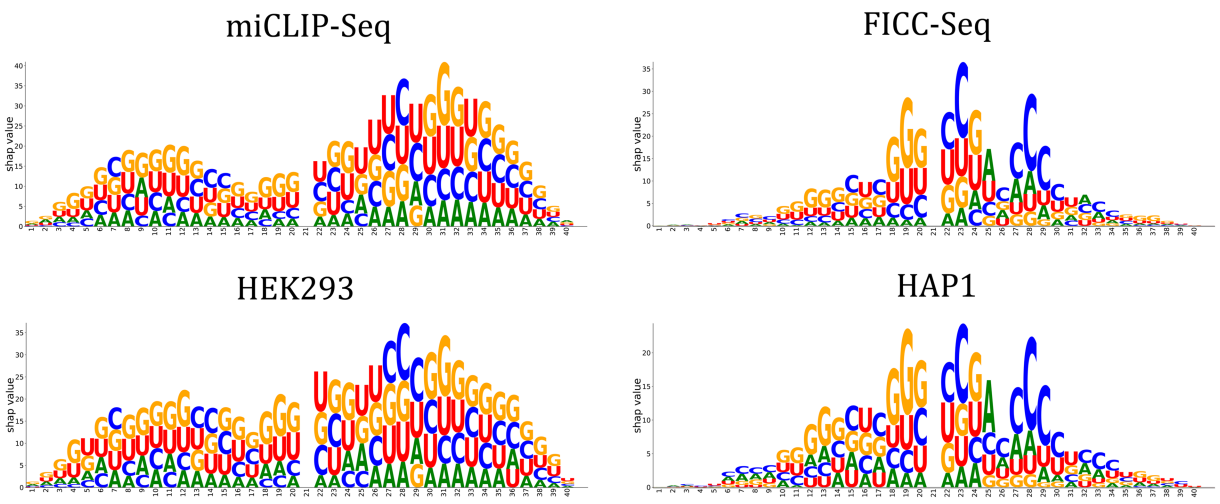


FIGURE 4
Factors affecting m5U identification for different datasets are revealed by DeepSHAP.

the Recall scores ranged from 39.44 to 93.90%, with a mean value of 61.89% (Table 2; Supplementary Figure S6). Closer examination of the inter-dataset evaluation revealed one interesting observation. Deepm5U models trained on miCLIP-Seq dataset or HEK293 dataset can better predict FICC-Seq dataset or HAP1 dataset, but not vice versa. The Recall score of the former (miCLIP-Seq Recall: 89.14%, HEK293 Recall: 93.90%) was nearly twice as high as that of the latter (FICC-Seq Recall: 48.41%, HAP1 Recall: 44.77%), especially for the mature mRNA mode. This phenomenon can be well explained by our DeepSHAP method (Figure 4). The distribution of SHAP values for these datasets was significantly different. The features near the central m^5U sites played the most important role in the FICC-Seq dataset and HAP1 dataset, while for both miCLIP-Seq dataset and HEK293 dataset, features at many positions contributed to the predictions. In terms of the total SHAP values, the FICC-Seq and HAP1 datasets did not contain enough feature information to support their accurate predictions for the other two datasets.

Variant predicting probabilities from saturation mutagenesis reveals potential valuable region

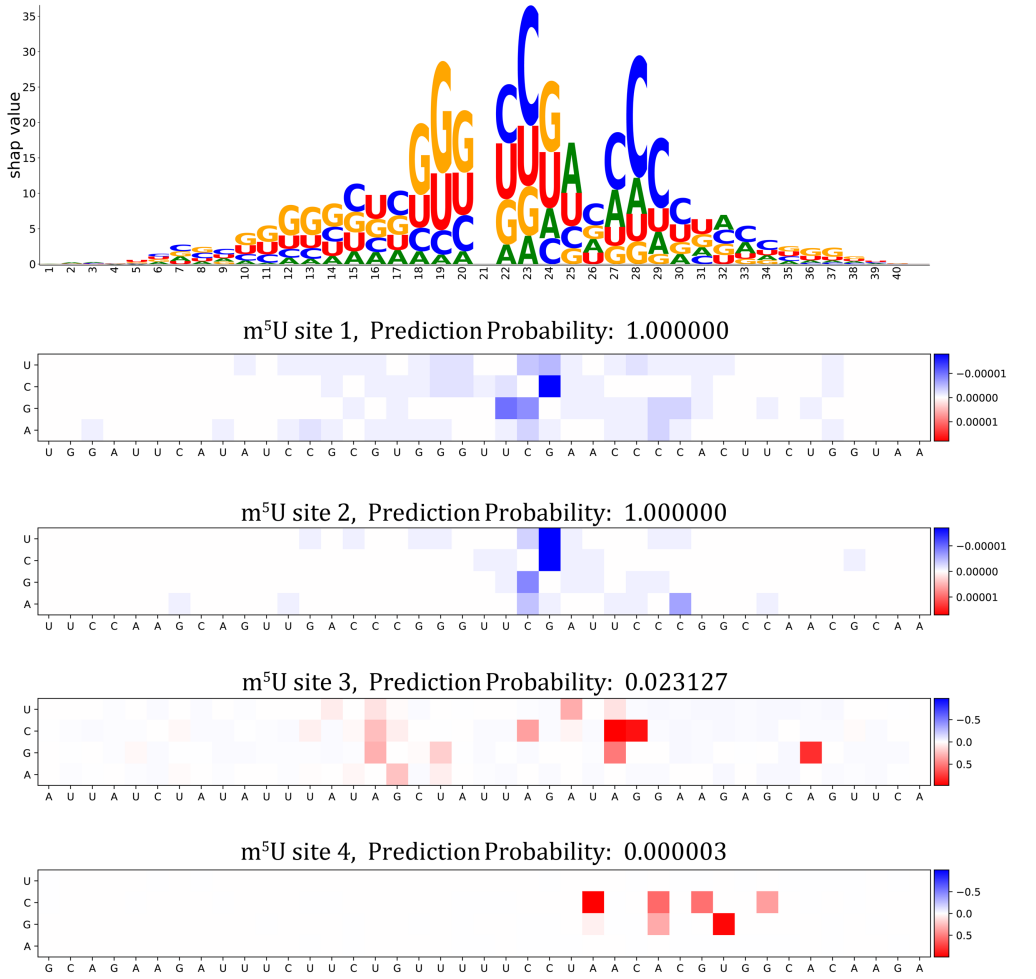
autoBioSeqpy supports variant effect prediction (Figure 5). We first called the instruction 5 to train the CNN-BiLSTM model on the FICC-Seq dataset. After that, we performed *in silico* saturation mutagenesis of four experimentally verified m^5U sites, using instruction 6 to convert every position in the sequence to every other possible base. We predicted these mutated sequences (instruction 7) and calculated mutation effect scores by measuring the changes in their predicted probabilities (instruction 8). A Jupyter notebook tutorial was provided for showing the details of plot generation.[2]

According to the distribution displayed in the heatmap, we found that for the sites which the CNN-BiLSTM model correctly predicted, the difference values were very small (less than 1e-4), but the sites which did not correctly predicted by the CNN-BiLSTM model, a part of the mutation increased the probability by more than 50%. This observation can be explained by the structure of LSTM structure that few change of a word vector (i.e., the mixture one-hot encoded bases by CNN in this work) will change the hidden states and thus will affect the final decision layer. Based on this property, the region contain high different probabilities can be a reference for further research. At the same time, we also observed that the distribution of high difference probabilities in heat map is similar with the region of large SHAP values, which could support that the region of high difference probabilities contains research value.

Performance comparison of Deepm5U with the exiting method

We compared Deepm5U's performance against the recently published algorithm m5UPred, which was that trained and tested on the same benchmark datasets as ours. Deepm5U produced more accurate training classification results (*ACC* scores of 92.32 and 91.47%) for full transcript mode and mature mRNA mode, respectively, than does m5UPred (88.32%, 89.91%; Table 3). We also observed better independent test performance with Deepm5U (*ACC* scores of 92.91% and 92.48%% for full transcript mode and mature mRNA mode) than that of m5UPred (88.35%, 89.70%). For the cross-technique and cross-cell-type evaluations, the comparison results were shown in Supplementary Table S7. In this comparison we see that Deepm5U produced the better performance (miCLIP-Seq *MCC*: 0.7093 and 0.8945, FICC-Seq *MCC*: 0.5574 and 0.6574, HEK293 *MCC*: 0.8135 and 0.9181, HAP1 *MCC*: 0.5967 and 0.6408) compared with m5UPred (miCLIP-Seq *MCC*: 0.6520 and 0.8090, FICC-Seq *MCC*: 0.4950 and 0.4490, HEK293 *MCC*: 0.7260 and 0.8450, HAP1 *MCC*: 0.5070 and 0.4610) for the two modes.

2 https://github.com/jingry/autoBioSeqpy/blob/2.0/notebook/An%20Example%20of%20Mutation%20Plotting.ipynb



Instruction 5:

```
python running.py --dataType rna --dataEncodingType onehot --dataTrainFilePaths data/FICC-Seq_Full_positive.txt data/FICC-Seq_Full_negative.txt --dataTrainLabel 1 0 --dataSplitScale 0.8 --modelLoadFile model/CNN-BiLSTM.py --verbose 1 --outSaveFolderPath tmpOut --savePrediction 1 --saveFig 1 --batch_size 64 --epochs 20 --shuffleDataTrain 1 --spcLen 41 --modelSaveName tmpMod.json --weightSaveName tmpWeight.bin --noGPU 0 --paraSaveName parameters.txt optimizer optimizers.Adam(lr=0.001,amsgrad=False,decay=False)
```

Instruction 6:

```
python tool/generateMutatedSeqs.py --outFile tmpOut/mutatedSeqs.txt --withOri 1 --oriName tmp --randomly 1 --oriSeq UGGAUUCAUAUCCGCGUGGGUUCGAACCCCACUUCUGGUAA --dataType rna
```

Instruction 7:

```
Python predicting.py --paraFile tmpOut/parameters.txt –dataTestFilePaths tmpOut/mutatedSeqs.txt --predictionSavePath tmpOut/indPredictions.txt
```

Instruction 8:

```
python tool/findDiff.py tmpOut/indPredictions.txt > tmpOut/diffMat.txt
```

FIGURE 5
autoBioSeqpy visualization of *in silico* mutagenesis on the CNN-BiLSTM model for four selected m^{5}U sequences in the FICC-Seq dataset.

Discussion

RNA chemical modifications can influence biological function. Accurate transcriptome-wide mapping and single-nucleotide resolution detection of these dynamic RNA modifications are critical for understanding gene regulation and function. In recent years, deep learning methods have provided remarkably good results in various biological applications, including the identification of various epitranscriptomic marks. Nevertheless, choosing the best-suited models and proper fine-tuning strategies remains a significant challenge for the development of personalized prediction algorithms based on the user's data. There is a pressing need to develop

TABLE 3 Performance comparison of Deepm5U and m5UPred on the training and independent test datasets for two modes.

Method	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)	MCC	auROC	auPR
Full transcript (Training dataset)							
m5UPred	88.32	–	87.59	–	0.7670	0.9560	–
Deepm5U	92.32	92.29	91.93	92.65	0.8465	0.9740	0.9781
Full transcript (Independent test set)							
m5UPred	88.35	–	87.90	–	0.7670	0.9560	–
Deepm5U	92.91	92.96	93.79	92.16	0.8584	0.9773	0.9810
Mature mRNA (Training dataset)							
m5UPred	89.91	–	88.64	–	0.7980	0.9560	–
Deepm5U	91.47	91.04	89.83	92.58	0.8304	0.9640	0.9648
Mature mRNA (Independent test set)							
m5UPred	89.70	–	87.44	–	0.7950	0.9540	–
Deepm5U	92.48	92.47	92.65	92.30	0.8498	0.9511	0.9473

user-friendly and model-adjustable environments for building and running DL models.

autoBioSeqpy is our contribution to the field for the accessibility and dissemination of deep learning techniques in biology. The autoBioSeqpy environment facilitates the creation of reproducible workflows and results for developers and end users, reducing the tedious modeling process in the routinely performed biological sequence classification tasks. By leveraging autoBioSeqpy, here we have explored the use of DL methods to identify RNA modifications such as m^5U methylation. Various common DL model architectures were evaluated, including CNN, BiGRU, BiLSTM, CNN-BiGRU and CNN-BiLSTM. Our systematic and comprehensive benchmark study suggests that deep-learning-based algorithms that rely only on RNA sequence are effective in predicting potential m^5U sites, outperforming current state-of-the-art tool. In particular, the performance of CNN-BiLSTM model was consistently better than all other DL models and was therefore chosen as the final predictor, called Deepm5U, for subsequent experiments and comparisons. We have also introduced two interpretability approaches to elucidate the mechanism of model and the influence of features. This has explained quite a few interesting phenomena that cannot be uncovered by conventional motif analysis. Furthermore, we have provided a step-by-step guide on how to use autoBioSeqpy to run model development and analysis tasks, and hope that this strategy can be extended to facilitate the study of other RNA modifications.

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

LY designed the study, collected the samples, and wrote the manuscript. YZ performed the experiments and analyzed the data. LX and FL analyzed the data. RJ designed the study and performed the experiments. JL designed the study, analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 22203057) Fund of Science and Technology Department of Guizhou Province ([2017]5790–07), Natural Science Foundation of Department of Education of Guizhou Province ([2021]021), Joint project of Luzhou Municipal People's Government and Southwest Medical University (2020LZXNYDJ39).

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.2023.1175925/full#supplementary-material

References

Alam, W., Ali, S. D., Tayara, H., and Chong, K. T. (2020). A CNN-based RNA N6-Methyladenosine site predictor for multiple species using heterogeneous features representation. *IEEE Access* 8, 138203–138209. doi: 10.1109/ACCESS.2020.3002995

Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N., and Tavazoie, S. F. (2015). N6-methyladenosine marks primary microRNAs for processing. *Nature* 519, 482–485. doi: 10.1038/nature14281

Ali, S. D., Kim, J. H., Tayara, H., and Chong, K. T. (2021). Prediction of RNA 5-Hydroxymethylcytosine modifications using deep learning. *IEEE Access* 9, 8491–8496. doi: 10.1109/ACCESS.2021.3049146

Barbieri, I., and Kouzarides, T. (2020). Role of RNA modifications in cancer. *Nat. Rev. Cancer* 20, 303–322. doi: 10.1038/s41568-020-0253-2

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., et al. (2022). MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* 50, D231–D235. doi: 10.1093/nar/gkab1083

Carter, J. M., Emmett, W., Mozos, I. R., Kotter, A., Helm, M., Ule, J., et al. (2019). FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. *Nucleic Acids Res.* 47:e113. doi: 10.1093/nar/gkz658

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A. I., Webb, G. I., et al. (2020). Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief. Bioinform.* 21, 1676–1696. doi: 10.1093/bib/bbz112

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar: Association for Computational Linguistics. 103–111. doi: 10.3115/v1/w14-4012

Chollet, F. (2015). Keras: the Python deep learning API [WWW Document]. Available at: https://github.com/fchollet/keras

Delaunay, S., and Frye, M. (2019). RNA modifications regulating cell fate in cancer. *Nat. Cell Biol.* 21, 552–559. doi: 10.1038/s41556-019-0319-0

El Allali, A., Elhamraoui, Z., and Daoud, R. (2021). Machine learning applications in RNA modification sites prediction. *Comput. Struct. Biotechnol. J.* 19, 5510–5524. doi: 10.1016/j.csbj.2021.09.025

Feng, P., and Chen, W. (2022). iRNA-m5U: a sequence based predictor for identifying 5-methyluridine modification sites in *Saccharomyces cerevisiae*. *Methods* 203, 28–31. doi: 10.1016/j.ymeth.2021.04.013

Han, Y., Feng, J., Xia, L., Dong, X., Zhang, X., Zhang, S., et al. (2019). CVm6A: a visualization and exploration database for m^{6}As in cell lines. *Cells* 8:168. doi: 10.3390/cells8020168

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jiang, J., Song, B., Tang, Y., Chen, K., Wei, Z., and Meng, J. (2020). m5UPred: a web server for the prediction of RNA 5-Methyluridine sites from sequences. *Mol. Ther. Nucleic Acids* 22, 742–747. doi: 10.1016/j.omtn.2020.09.031

Jin, X. B., Yang, A., Su, T., Kong, J. L., and Bai, Y. (2021). Multi-Channel fusion classification method based on time-series data. *Sensors (Basel)* 21:4391. doi: 10.3390/s21134391

Jing, R., Li, Y., Xue, L., Liu, F., Li, M., and Luo, J. (2020). autoBioSeqpy: a deep learning tool for the classification of biological sequences. *J. Chem. Inf. Model.* 60, 3755–3764. doi: 10.1021/acs.jcim.0c00409

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., and Novoa, E. M. (2017). The RNA modification landscape in human disease. *RNA* 23, 1754–1769. doi: 10.1261/rna.063503.117

Keffer-Wilkes, L. C., Soon, E. F., and Kothe, U. (2020). The methyltransferase TrmA facilitates tRNA folding through interaction with its RNA-binding domain. *Nucleic Acids Res.* 48, 7981–7990. doi: 10.1093/nar/gkaa548

Khan, S. M., He, F., Wang, D., Chen, Y., and Xu, D. (2020). MU-PseUDeep: a deep learning method for prediction of pseudouridine sites. *Comput. Struct. Biotechnol. J.* 18, 1877–1883. doi: 10.1016/j.csbj.2020.07.010

Kim, N., Kim, H. K., Lee, S., Seo, J. H., Choi, J. W., Park, J., et al. (2020). Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* 38, 1328–1336. doi: 10.1038/s41587-020-0537-9

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, H., Chen, L., Huang, Z., Luo, X., Li, H., Ren, J., et al. (2021). DeepOMe: a web server for the prediction of 2'-O-me sites based on the hybrid CNN and BLSTM architecture. *Front. Cell Dev. Biol.* 9:686894. doi: 10.3389/fcell.2021.686894

Li, Z., Mao, J., Huang, D., Song, B., and Meng, J. (2022). RNADSN: transfer-learning 5-Methyluridine (m5U) modification on mRNAs from common features of tRNA. *Int. J. Mol. Sci.* 23:13493. doi: 10.3390/ijms232113493

Li, X., Xiong, X., and Yi, C. (2016). Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat. Methods* 14, 23–31. doi: 10.1038/nmeth.4110

Liu, H., Flores, M. A., Meng, J., Zhang, L., Zhao, X., Rao, M. K., et al. (2015). MeT-DB: a database of transcriptome methylation in mammalian cells. *Nucleic Acids Res.* 43, D197–D203. doi: 10.1093/nar/gku1024

Liu, H., Ma, J., Meng, J., and Zhang, L. (2021). MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine Methyltranscriptome. *Methods Mol. Biol.* 2284, 507–518. doi: 10.1007/978-1-0716-1307-8_27

Liu, H., Wang, H., Wei, Z., Zhang, S., Hua, G., Zhang, S. W., et al. (2018). MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res.* 46, D281–D287. doi: 10.1093/nar/gkx1080

Liu, S., Zhu, A., He, C., and Chen, M. (2020). REPIC: a database for exploring the N6-methyladenosine methylome. *Genome Biol.* 21:100. doi: 10.1186/s13059-020-02012-4

Luo, X., Li, H., Liang, J., Zhao, Q., Xie, Y., Ren, J., et al. (2021). RMVar: an updated database of functional variants involved in RNA modifications. *Nucleic Acids Res.* 49, D1405–D1412. doi: 10.1093/nar/gkaa811

Ma, J., Song, B., Wei, Z., Huang, D., Zhang, Y., Su, J., et al. (2022). m5C-atlas: a comprehensive database for decoding and annotating the 5-methylcytosine (m5C) epitranscriptome. *Nucleic Acids Res.* 50, D196–D203. doi: 10.1093/nar/gkab1075

McInnes, L., and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. [Preprint]. https://arxiv.org/abs/1802.03426

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., et al. (2015). 5' UTR m(6)a promotes cap-independent translation. *Cells* 163, 999–1010. doi: 10.1016/j.cell.2015.10.012

Mostavi, M., Salekin, S., and Huang, Y. (2018). Deep-2'-O-me: predicting 2'-O-methylation sites by convolutional neural networks. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 2394–2397. doi: 10.1109/EMBC.2018.8512780

Nachtergaele, S., and He, C. (2018). Chemical modifications in the life of an mRNA transcript. *Annu. Rev. Genet.* 52, 349–372. doi: 10.1146/annurev-genet-120417-031522

Nordlund, M. E., Johansson, J. O., von Pawel-Rammingen, U., and Byström, A. S. (2000). Identification of the TRM2 gene encoding the tRNA (m5U54) methyltransferase of *Saccharomyces cerevisiae*. *RNA* 6, 844–860. doi: 10.1017/s1355838200992422

Pereira, M., Ribeiro, D. R., Pinheiro, M. M., Ferreira, M., Kellner, S., and Soares, A. R. (2021). m5U54 tRNA Hypomodification by lack of TRMT2A drives the generation of tRNA-derived small RNAs. *Int. J. Mol. Sci.* 22:2941. doi: 10.3390/ijms22062941

Phizicky, E. M., and Alfonzo, J. D. (2010). Do all modifications benefit all tRNAs? *FEBS Lett.* 584, 265–271. doi: 10.1016/j.febslet.2009.11.049

Powell, C. A., and Minczuk, M. (2020). TRMT2B is responsible for both tRNA and rRNA m5U-methylation in human mitochondria. *RNA Biol.* 17, 451–462. doi: 10.1080/15476286.2020.1712544

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cells* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Sarkar, A., Gasperi, W., Begley, U., Nevins, S., Huber, S. M., Dedon, P. C., et al. (2021). Detecting the epitranscriptome. *Wiley Interdiscip. Rev. RNA* 12:e1663. doi: 10.1002/wrna.1663

Song, B., Tang, Y., Chen, K., Wei, Z., Rong, R., Lu, Z., et al. (2020). m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. *Bioinformatics* 36, 3528–3536. doi: 10.1093/bioinformatics/btaa178

Su, D., Chan, C. T. Y., Gu, C., Lim, K. S., Chionh, Y. H., McBee, M. E., et al. (2014). Quantitative analysis of ribonucleoside modifications in tRNA by HPLC-coupled mass spectrometry. *Nat. Protoc.* 9, 828–841. doi: 10.1038/nprot.2014.047

Sun, W. J., Li, J. H., Liu, S., Wu, J., Zhou, H., Qu, L. H., et al. (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.* 44, D259–D265. doi: 10.1093/nar/gkv1036

Suzuki, T., Ueda, H., Okada, S., and Sakurai, M. (2015). Transcriptome-wide identification of adenosine-to-inosine editing using the ICE-seq method. *Nat. Protoc.* 10, 715–732. doi: 10.1038/nprot.2015.037

Tahir, M., Tayara, H., and Chong, K. T. (2019a). iPseU-CNN: identifying RNA Pseudouridine sites using convolutional neural networks. *Mol. Ther. Nucleic Acids* 16, 463–470. doi: 10.1016/j.omtn.2019.03.010

Tahir, M., Tayara, H., and Chong, K. T. (2019b). iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components. *J. Theor. Biol.* 465, 1–6. doi: 10.1016/j.jtbi.2018.12.034

Tang, Y., Chen, K., Song, B., Ma, J., Wu, X., Xu, Q., et al. (2021). m6A-atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res.* 49, D134–D143. doi: 10.1093/nar/gkaa692

Tareen, A., and Kinney, J. B. (2020). Logomaker: beautiful sequence logos in Python. *Bioinformatics* 36, 2272–2274. doi: 10.1093/bioinformatics/btz921

Urbonavicius, J., Jäger, G., and Björk, G. R. (2007). Amino acid residues of the *Escherichia coli* tRNA (m5U54) methyltransferase (TrmA) critical for stability, covalent binding of tRNA and enzymatic activity. *Nucleic Acids Res.* 35, 3297–3305. doi: 10.1093/nar/gkm205

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Wang, C., Ju, Y., Zou, Q., and Lin, C. (2021). DeepAc4C: a convolutional neural network model with hybrid features composed of physicochemical patterns and distributed representation information for identification of N4-acetylcytidine in mRNA. *Bioinformatics* 38, 52–57. doi: 10.1093/bioinformatics/btab611

Wetzel, C., and Limbach, P. A. (2016). Mass spectrometry of modified RNAs: recent developments. *Analyst* 141, 16–23. doi: 10.1039/c5an01797a

Wiener, D., and Schwartz, S. (2021). The epitranscriptome beyond m^6A. *Nat. Rev. Genet.* 22, 119–131. doi: 10.1038/s41576-020-00295-8

Xuan, J. J., Sun, W. J., Lin, P. H., Zhou, K. R., Liu, S., Zheng, L. L., et al. (2018). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* 46, D327–D334. doi: 10.1093/nar/gkx934

Zhang, L., Chen, J., Ma, J., and Liu, H. (2021a). HN-CNN: a heterogeneous network based on convolutional neural network for m7 G site disease association prediction. *Front. Genet.* 12:655284. doi: 10.3389/fgene.2021.655284

Zhang, Y., and Hamada, M. (2018). DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics* 19:524. doi: 10.1186/s12859-018-2516-4

Zhang, Y., Jiang, J., Ma, J., Wei, Z., Wang, Y., Song, B., et al. (2023). DirectRMDB: a database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology. *Nucleic Acids Res.* 51, D106–D116. doi: 10.1093/nar/gkac1061

Zhang, G., Luo, W., Lyu, J., Yu, Z. G., and Huang, G. (2022). CNNLSTMac4CPred: a hybrid model for N4-Acetylcytidine prediction. *Interdiscip. Sci.* 14, 439–451. doi: 10.1007/s12539-021-00500-0

Zhang, L., Qin, X., Liu, M., Xu, Z., and Liu, G. (2021b). DNN-m6A: a cross-species method for identifying RNA N6-Methyladenosine sites based on deep neural network with multi-information fusion. *Genes* 12:354. doi: 10.3390/genes12030354

Zheng, Y., Nie, P., Peng, D., He, Z., Liu, M., Xie, Y., et al. (2018). m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res.* 46, D139–D145. doi: 10.1093/nar/gkx895

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Zhuang, J., Liu, D., Lin, M., Qiu, W., Liu, J., and Chen, S. (2021). PseUdeep: RNA Pseudouridine site identification with deep learning algorithm. *Front. Genet.* 12:773882. doi: 10.3389/fgene.2021.773882

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 19 July 2023
DOI 10.3389/fmicb.2023.1226166

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
J. Christopher Fenno,
University of Michigan, United States
Miroslaw Ksiazek,
Jagiellonian University, Poland

*CORRESPONDENCE
Cristian Cortez
✉ cristian.cortez@pucv.cl
Juan P. Cárdenas
✉ juan.cardenas@umayor.cl

†These authors have contributed equally to this work

RECEIVED 20 May 2023
ACCEPTED 04 July 2023
PUBLISHED 19 July 2023

CITATION
Morales-Olavarría M, Nuñez-Belmar J, González D, Vicencio E, Rivas-Pardo JA, Cortez C and Cárdenas JP (2023) Phylogenomic analysis of the *Porphyromonas gingivalis - Porphyromonas gulae* duo: approaches to the origin of periodontitis. *Front. Microbiol.* 14:1226166. doi: 10.3389/fmicb.2023.1226166

COPYRIGHT
© 2023 Morales-Olavarría, Nuñez-Belmar, González, Vicencio, Rivas-Pardo, Cortez and Cárdenas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Phylogenomic analysis of the *Porphyromonas gingivalis - Porphyromonas gulae* duo: approaches to the origin of periodontitis

Mauricio Morales-Olavarría[1†], Josefa Nuñez-Belmar[1†], Dámariz González[1], Emiliano Vicencio[2], Jaime Andres Rivas-Pardo[1,3], Cristian Cortez[2*] and Juan P. Cárdenas[1,3*]

[1]Centro de Genómica y Bioinformática, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Santiago, Chile, [2]Escuela de Tecnología Médica, Facultad de Ciencias, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, [3]Escuela de Biotecnología, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Santiago, Chile

Porphyromonas gingivalis is an oral human pathogen associated with the onset and progression of periodontitis, a chronic immune-inflammatory disease characterized by the destruction of the teeth-supporting tissue. *P. gingivalis* belongs to the genus *Porphyromonas,* which is characterized by being composed of Gram-negative, asaccharolytic, non-spore-forming, non-motile, obligatory anaerobic species, inhabiting niches such as the oral cavity, urogenital tract, gastrointestinal tract and infected wound from different mammals including humans. Among the *Porphyromonas* genus, *P. gingivalis* stands out for its specificity in colonizing the human oral cavity and its keystone pathogen role in periodontitis pathogenesis. To understand the evolutionary process behind *P. gingivalis* in the context of the *Pophyoromonas* genus, in this study, we performed a comparative genomics study with publicly available *Porphyromonas* genomes, focused on four main objectives: (A) to confirm the phylogenetic position of *P. gingivalis* in the *Porphyromonas* genus by phylogenomic analysis; (B) the definition and comparison of the pangenomes of *P. gingivalis* and its relative *P. gulae*; and (C) the evaluation of the gene family gain/loss events during the divergence of *P. gingivalis* and *P. gulae*; (D) the evaluation of the evolutionary pressure (represented by the calculation of Tajima-D values and dN/dS ratios) comparing gene families of *P. gingivalis* and *P. gulae*. Our analysis found 84 high-quality assemblies representing *P. gingivalis* and 14 *P. gulae* strains (from a total of 233 *Porphyromonas* genomes). Phylogenomic analysis confirmed that *P. gingivalis* and *P. gulae* are highly related lineages, close to *P. loveana*. Both organisms harbored open pangenomes, with a strong core-to-accessory ratio for housekeeping genes and a negative ratio for unknown function genes. Our analyses also characterized the gene set differentiating *P. gulae* from *P. gingivalis*, mainly associated with unknown functions. Relevant virulence factors, such as the FimA, Mfa1, and the hemagglutinins, are conserved in *P. gulae, P. gingivalis,* and *P. loveana*, suggesting that the origin of those factors occurred previous to the *P. gulae - P. gingivalis* divergence. These results suggest an unexpected evolutionary relationship between the *P. gulae - P. gingivalis* duo and *P. loveana*, showing more clues about the origin of the role of those organisms in periodontitis.

KEYWORDS

Porphyromonas gingivalis, virulence factors, phylogenomics, orthogroups, dN/dS, Tajima D value, gene gain/loss model

1. Introduction

The *Porphyromonas* genus comprises Gram-negative, asaccharolytic, non-spore-forming, non-motile, and obligate anaerobic species (Paster et al., 1994; Summanen and Finegold, 2015). Its ecological niche is mainly the oral cavity of humans and other mammals, therefore being described by some authors as indigenous to the oral microbiome. It has also been isolated from the urogenital and gastrointestinal tracts and infected wound sites in humans and other mammals (Summanen et al., 2009; Sakamoto and Ohkuma, 2013; Rajilić-Stojanović and de Vos, 2014; Sakamoto et al., 2015; Sato et al., 2015; Zamora-Cintas et al., 2018; Guilloux et al., 2021). There are several officially described *Porphyromonas* species (Gibson and Genco, 2006). Of these species, *Porphyromonas gingivalis* is one of the most studied because of its key role in the pathogenesis of human periodontitis. It is a chronic, non-communicable, immuno-inflammatory disease characterized by the irreversible destruction of dental supporting tissues, collectively known as the periodontium. Periodontitis is caused by the constant challenge of the dysbiotic subgingival biofilm, where *P. gingivalis* acts as a keystone pathogen. Therefore, the uncontrolled immune response against these pathogens is responsible for the destructive nature of this disease (Hajishengallis and Lamont, 2014).

Due to its prevalent preference for the oral cavity, *Porphyromonas* evolution could be influenced by niche construction, a process in which the organisms periodically change the local resource distribution in ways that enhance their own fitness. As a keystone pathogen, the large impact on the structure and function of microbial communities can be influenced by perturbations in the microenvironment caused by niche construction, with a concomitant effect on the evolution of itself and the community (Laland et al., 1999). For example, *P. gingivalis* is closely related to *P. gulae*, a member of the oral microbiome of various mammalian hosts (Fournier et al., 2001); additionally, some virulence factors found in *P. gingivalis*, such as the fimbrial proteins, are also present in *P. gulae,* showing high sequence conservation (Hamada et al., 2008; Fujiwara-Takahashi et al., 2020).

P. gingivalis, as a key member of the dysbiotic subgingival microbiota, is capable to colonize, invade, and destroy tooth-supporting tissues (periodontium) via the expression of several virulence factors (Hajishengallis and Lamont, 2014; Zenobia and Hajishengallis, 2015). Despite having a relatively low abundance in the subgingival biofilm, this pathogen has keystone abilities that cause microbial dysbiosis and alter the host's immune response, promoting the pathologic destructive inflammatory environment and potentially influencing the course of other inflammatory disorders, even causing systemic complications (Olsen et al., 2017).

This pathogen synthesizes, expresses, and secretes several virulence factors contributing to its pathogenesis, including lipopolysaccharide (LPS), gingipains, hemagglutinins, and fimbriae, among others (Nakayama and Ohara, 2017). All those factors contribute to the damage of periodontal tissue and exacerbate inflammatory responses. For instance, LPS released after bacterial lysis and secreted through outer membrane vesicles (OMVs) activates Toll-like receptor 4 (TLR4) (Waller et al., 2016), promoting destructive periodontal inflammation (Zhang D. et al., 2015). The cysteine protease family members, the gingipains, play crucial roles in complex host-pathogen interactions. In multispecies heterotypic communities, its protease activity degrades ligands linked to interspecific aggregation and coaggregation, influencing the structure and composition of biofilms (Jia et al., 2019). Moreover, the degradation of host proteins as a pathogenicity strategy enables the bacterium to evade the immune response (Chen et al., 2023), modulate signaling pathways that increase its virulence (Lourbakos et al., 2001), and acquire essential nutrients during infection (Guo et al., 2010). Additionally, fimbriae play a key role in host tissue adhesion (Chen et al., 2023) as well as bacterial aggregation and coaggregation, two processes necessary for the formation of oral biofilms (Hasegawa and Nagano, 2021). The two main fimbriae systems in *P. gingivalis,* FimA, and Mfa1, are involved in the interaction with the host-epithelial surface, the immune response, autoaggregation, and the aggregation of other oral bacteria, such as *Treponema* spp. and *Streptococcus gordonii* (Chung et al., 2000; Yamada et al., 2005; Hajishengallis et al., 2006; Harokopakis et al., 2006; Eskan et al., 2007; Kuboniwa et al., 2009). The *fim* and *mfa* gene clusters contain a diversity of polymorphisms that could be associated with different genotypes and are often utilized to classify *P. gingivalis* strains (Hasegawa and Nagano, 2021). In the case of periodontitis-like diseases in other mammals, *P. gulae* adhesion systems have been reported as key players in the development of periodontitis; moreover, *P. gulae* can invade human gingival epithelial cells *in vitro*, and its efficiency could be associated with the *fimA* genotype (Inaba et al., 2019). In cats, types B and C fimbriae were frequently detected in subjects with periodontitis, which increased the use of *fimA* as a marker related to virulences in periodontitis (Iwashita et al., 2019).

The current knowledge available indicates that there are several *Porphyromonas* species and that only a few species are involved in periodontal disease, being *P. gingivalis* and *P. gulae*. The availability of genomic information for different *Porphyromonas* strains and metagenome-assembled genomes (MAGs) raises the opportunity for the genomic analysis of several aspects of evolution and development of the pathogenic properties of *P. gingivalis* in the context of its genus and its close relatedness with *P. gulae*. Despite there are some genomic studies comparing *P. gingivalis* strains in order to make associations between genetic content and its pathogenic properties (Chen et al., 2004; Lenzo et al., 2016; Chen et al., 2017; Romero-Lastra et al., 2017; Mendez et al., 2019), currently, there are only a few studies covering possible relationships between evolution and gene content for *P. gingivalis* in comparison with other microorganisms. One remarkable study performed with 32 isolates from 18 species (O'Flynn et al., 2015) showed that *P. gingivalis* genomes contained some distinctive features, such as genes involved in iron transport and heme utilization, whereas lacked other genes (involved in processes such as

protoporphyrin biosynthesis), when they were compared with sequences from other *Porphyromonas* species. Another study (Endo et al., 2015) compared *P. gingivalis* genomes with the sequences from other members of the so-called "red complex" (*Treponema denticola* and *Tannerella forsythia*), a set of organisms involved in the development of periodontal disease (Mohanty et al., 2019); this study predicted metabolic complementations in the fatty acid biosynthesis pathways between those organisms, suggesting a potential effect of pathway complementation in the development of their roles in disease.

Despite the availability of previous studies and genomic information, the relationship between genetic diversity, pathogenesis, and evolution of the adaptation in *P. gingivalis* to its role in human periodontitis remains to be clarified. Considering the potential evolutive history of *P. gingivalis* in the context of its genus, and the existence of other species such as *P. gulae* exhibiting similar characteristics in terms of niche and mechanism of infection in other mammals, a comparative genomics study with an evolutive perspective can help to understand how *P. gingivalis* acquired its features to be a keystone pathogen. Since virulence is the product of complex pathogen-host interactions (Diard and Hardt, 2017); the evolutionary analysis of virulence factors could help to explain the acquirement of the lifestyle of a keystone pathogen, is an important milestone to understand the evolution of some members of *Porphyromonas*. In addition to virulence factors, other accessory proteins in the machinery of *Porphyromonas* could show evidence of adaptation to the oral microbiome. Understanding the evolutionary process behind the evolution of *P. gingivalis* as a keystone pathogen could give signs of how other *Porphyromonas* had evolved in a different way to potentially become a keystone pathogen, like potentially could be *P. gulae*. In this study, we describe *Porphyromonas* pangenome and explore the phylogenetic relation of *P. gingivalis* virulence with other members of the genus, especially *P. gulae*, the closest species to *P. gingivalis*.

2. Methodology

2.1. Genome dataset definition

Porphyromonas genomes were selected from NCBI Genbank FTP site (August 2022). The dataset was filtered using two criteria: their taxonomic genomic identity and degree of completeness and contamination. The taxonomic identity was established by using the program '*classify_wf*' of the GTDB-TK program, version 2.1.0 (Chaumeil et al., 2019), using the database release 207 as the reference, selecting all genomes classified into the *Porphyromonas* genus ("g__*Porphyromonas*" or "g__*Porphyromonas*_A"). Genome completeness and contamination were calculated using the program '*lineage_wf*' from CheckM version 1.1.3 (Parks et al., 2015); only those genomes with completeness equal to or higher than 90%, and contamination below 5%, were selected, as previously suggested for "high-quality drafts" (Bowers et al., 2017).

2.2. Definition of species groups

In order to detect the genomic species represented among the selected genomes, we combined the prediction from GTDB-TK (see above) with the prediction of clusters defined by average nucleotide identity (ANI) values. All genomes were compared in an all-*vs*-all manner using FastANI version 1.32 (Jain et al., 2018) with default parameters. The raw pairwise comparison data was filtered, discarding all ANI values below 95%, the classical intra-species boundary for microbial genomes (Richter and Rosselló-Móra, 2009). Filtered pairwise comparisons were analyzed by the *MCL* program, creating putative genomic species clusters as observed in network clustering (van Dongen and Abreu-Goodger, 2012).

2.3. Annotation, identification of orthogroups, and phylogenomic tree

All members of the final dataset were annotated *de novo* using Prokka, version 1.11 (Seemann, 2014) (relevant parameters: *-metagenome -kingdom Bacteria -addgenes*). Orthogroups from the set of the final dataset were calculated by Orthofinder version 2.5.5 (Emms and Kelly, 2019) with the '*-og*' parameter. For phylogenomic tree elaboration, a concatenated multiple sequence alignment was constructed from a set of 38 single-copy conserved orthogroups by using MAFFT v. 7.490 (parameters: *-maxiterate 1,000 -localpair*) (Katoh et al., 2019). The alignment was used by *iqtree* v. 2.1.4 (Nguyen et al., 2015) (parameters: *-m TEST -alrt 1,000*) to generate a maximum likelihood-based tree with an aLRT with 1,000 replicates as the branch support test. The phylogenomic tree was visualized using the *Toytree* python package (Eaton, 2020), or with the tool FigTree v.1.4.4[1] when it corresponds. The genomic information from *Tannerella forsythia* (assembly GCA_000238215.1) was utilized as the outgroup for the *Porphyomonas* genus tree due to its relatedness with the genus (Summanen and Finegold, 2015).

2.4. Pangenome analyses and definition

The pangenome represents the collection of all groups of orthologous genes (orthogroups) from a set of genomes (Moldovan and Gelfand, 2018). To analyze the pangenome of *P. gingivalis* and *P. gulae*, separated Orthofinder executions were performed with the proteomes from each cluster without any outgroup (parameters: *-M msa -y*). The orthogroup matrix (including unassigned orthogroups) was obtained for each run and utilized for different pangenome metrics. Pangenome curves were created using the *panplots* function in R (created by *SioStef*),[2] using 1,000 permutations. The *alpha* values for Heap's law equation were calculated using the *curve_fit* function of the *scipy* package in Python, using as objective the equation "*a * (x ** b)*." The determination of shell, cloud, "soft-core" and core components of the pangenome was deduced from the complete orthogroup matrix by using Python scripts with the *pandas* package, considering *core* gene families as the orthogroups present in 100% of the strains, soft-core groups as present in between 90 and 99.9% of the strains, shell as groups present between 89 and 15% of strains, and cloud as the gene families present in between 14% and the equivalent to two strains. Unique groups can be deduced from the set of "species-specific orthogroups," and the "unassigned genes," both reported by

1 https://github.com/rambaut/figtree

2 https://github.com/SioStef/panplots/

Orthofinder. Figures were created with *ggplot2* and the *ggarrangment* R packages.

2.5. Predictive proteomic analysis

For predicted proteomes was made by combining the prediction of EggNOG mapper version 2.1.6 (Cantalapiedra et al., 2021) (parameters: "--tax_scope_mode narrowest --tax_scope prokaryota_broad --go_evidence experimental").

2.6. Gene gain/loss model for *Porphyromonas gingivalis* - *Porphyromonas gulae* divergence tree

In order to create a general gene gain/loss model, all proteomes from genomes from *P. gingivalis* and *P. gulae*, in addition to one genome of *P. loveana* as the outgroup were compared using Orthofinder as previously mentioned. The generated orthogroup matrix was utilized for the generation of a phylogenomic tree. This tree and the binary version of the complete orthogroup matrix were used by the software Count (Csurös, 2010) for the calculation of gene gain/loss rates following the Wagner parsimony model, using convergence criteria were set to a likelihood delta of 0.05 with a maximum of 1,000 rounds, and the same penalty score (equal to 1) for gains and losses. The final tree was represented using the ETE3 Python package (Huerta-Cepas et al., 2016).

2.7. Evolutionary metrics: Tajima D pairwise dN/dS ratios and selection among sites

Tajima's D value is a statistical test that infers rates of rare alleles and assigns scores to the orthogroups, therefore, detecting variation from a neutral model of molecular evolution (Zhang Q. et al., 2015). In order to calculate Tajima's D value, the nucleotide sequences from the coding sequences of the single-copy conserved orthogroups of *P. gingivalis*-only or *P. gulae*-only, when corresponding, were utilized. Sequences were aligned with MAFFT, as mentioned above. Nucleotide multiple alignments were utilized to calculate Tajima's D values using the *tajima.test* function from *pegas* R package.[3] Graphs were created with *ggplot2*.

Since that dN/dS is more suitable for comparisons between sequences with considerable distances (Rocha et al., 2006), dN/dS values were computed from the nucleotide sequence of the single-copy orthogroups conserved in *both P. gingivalis and P. gulae* genomes, using *P. loveana* as outgroup for both comparisons. Each orthogroup set was aligned using MAFFT (G-INS-i mode); calculation of dN, dS, and ω was made by using CODEML program from the PAML package(Yang, 2007), using the following parameters: "*runmode = −2, seqtype = 1, CodonFreq = 3, model = 1, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, omega = 0.5*." Pairwise comparisons with distances equal to zero, dN/dS > 5, and dS > 10 were discarded from the analysis. For all pairwise comparisons, the statistical significance among groups was evaluated using the Kolmogorov–Smirnov (*ks.test* in R) test (value of $p < 0.01$).

2.8. Phylogenetic analyses of virulence factors of *Porphyromonas*

Hidden Markov Model (HMM) profile search (using HMMER 3.1b) were performed for the following Pfam entries to search for the following virulence factors: PF06321 ("*P_gingi_FimA*," detectable in both FimA and FimC proteins), PF15495 ("*Major fimbrial subunit protein type IV, Fimbrillin, C-terminal*," found in Mfa1 proteins), PF07675 ("*Cleaved Adhesin Domain*"; present in both gingipains and hemagglutinins), PF10365 (*DUF2436*, as a confirmatory hit for gingipains and hemagglutinins), PF01364 ("*Peptidase family C25*," found in gingipains but not in hemagglutinins), and PF04371 ("*Porphyromonas-type peptidyl-arginine deiminase*"), found in the peptidyl-arginine deiminase (PAD) enzyme associated with protein citrullination in *P. gingivalis* (Gabarrini et al., 2015). Proteins from the *Porphyromonas* dataset were retrieved and aligned with MAFFT (G-INS-i mode). Phylogenetic trees of each PFAM were inferred using *iqtree* v. 2.1.4 (Nguyen et al., 2015) with the maximum likelihood (ML) method with an aLRT with 1,000 replicates as the branch support test (main parameters: *-m TEST --alrt 1,000*). For tree visualization, the FigTree v.1.4.4 software (see text footnote 1) was utilized.

3. Results

3.1. An up-to-date phylogeny of the *Porphyromonas* genus and representation of species groups

The *Porphyromonas* genus comprises a set of several organisms from both human and animal sources (Summanen and Finegold, 2015). Using the content available in the NCBI Genbank Genome repository, with the additional filter using taxonomic and completeness/contamination standards, the *Porphyromonas* dataset utilized in this study comprises 233 genomes, representing 36 genomic species clusters (Figure 1; Supplementary Figure S1; Supplementary Tables S1, S2). The five most represented genomic species in the dataset were *P. gingivalis*, *P. pasteri*, *P. levii*, *P. gulae*, *P. macacae* and *P. cangingivalis*, with 84, 34, 20, 14, 7, and 7 genomes, respectively. Of the set of genomic species, 19 clusters do not have an official species name; in most of those cases, only the GTDB-TK assignment was shown. Moreover, some strains, such as *Porphyromonas* sp. SUG530 (GCA_022772085.1), uncultured *Porphyromonas* sp. SRR2034640 _bin.3_metaWRAP_v1.1_MAG (GCA_915070105.1) or uncultured *Porphyromonas* sp. RxaearOaYr_bin.43.MAG (GCA_943912835.1), do not contain any species-level classification in GTDB-TK (Supplementary Table S1). The *P. uenonis* 60–3 strain (isolated from a vaginal sample) is part of a different species cluster than the canonical *P. uenonis* (DSM 23387, isolated from a sacral decubitus ulcer (Finegold et al., 2004)). This latter finding suggests that strain 60–3 could correspond to an undescribed species. The undescribed species (taxa with magenta names in Figure 1) cover an important fraction of the *Porphyromonas* tree,

3 https://cran.r-project.org/web/packages/pegas/index.html

10.3389/fmicb.2023.1226166

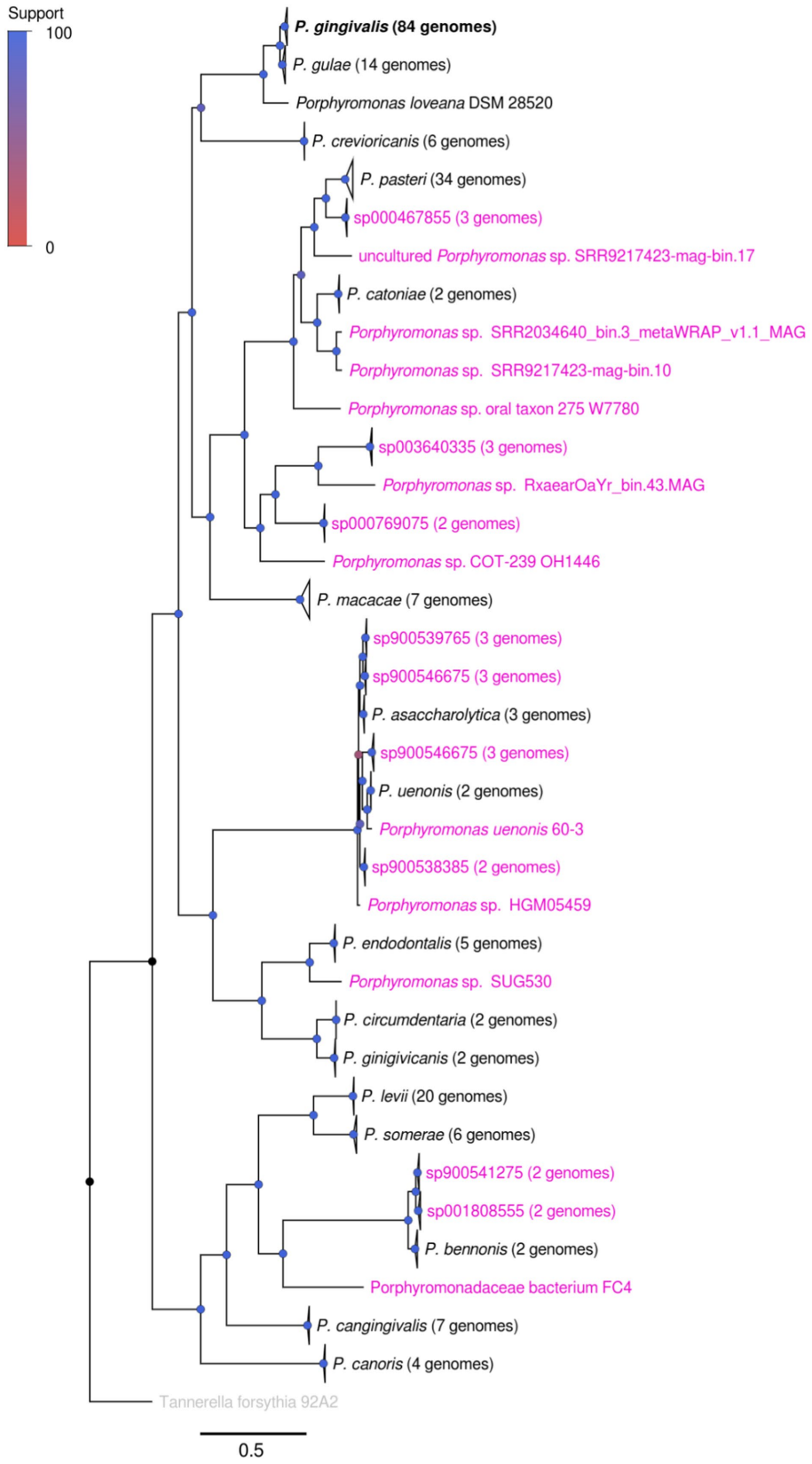


FIGURE 1
Phylogenomic tree showing the relationship between different *Porphyromonas* genomic species. The tree was created from the alignment of 38 single-copy conserved protein families, using the maximum likelihood method in IQTREE, with the use of the approximate likelihood ratio test (aLRT) as the branch support test. Taxa names colored in magenta represented genomes or species groups without an official name classification, according

(Continued)

FIGURE 1 (Continued)
to GTDB-TK and manual review (see Supplementary Table S2); taxa names in black represented species groups with an official name. Branch support is represented by colors in each node, according to the *Support* color scale.

suggesting an extensive uncharacterized genetic content among the genus.

This phylogenomic tree, made from 38 conserved orthogroups (Figure 1), confirmed several previously reported phylogenetic associations between different members of the *Porphyromonas* genus. For example, the closed relationship between *P. gingivalis* and *P. gulae* has been previously shown in different studies (Fournier et al., 2001). Interestingly, this study also found a strong relatedness between the *P. gulae - P. gingivalis* duo with P. *loveana*, an organism isolated from the oral cavity of a musky rat kangaroo, a marsupial (Bird et al., 2016). In addition to this finding, the close relationship between *P. uenonis* and *P. asaccharolytica* was also shown previously (Finegold et al., 2004), as well as the close relationship between *P. levii* and *P. somerae* (Summanen et al., 2009). However, from now on, this study will be focused on the genome set of 84 *P. gingivalis* and 14 *P. gulae* strains.

3.2. Comparing the pangenomes from *Porphyromonas gingivalis* and *Porphyromonas gulae*

Both the mutually highly related *P. gingivalis* and *P. gulae* are involved in periodontal diseases in their respective hosts (Lenzo et al., 2016). Therefore, studying evolutionary relationships between those organisms could be important to analyze to explain the origin of the *P. gingivalis* phenotype. Given that we found 84 high-quality genomes for *P. gingivalis* and 14 for *P. gulae*, we first compared them by analyzing each pangenome separately. For this purpose, the Orthofinder output was used to generate a gene matrix to calculate core, "persistent," shell, cloud, and unique gene families, as well as to calculate the pangenome accumulation curves (Figures 2A,B; Table 1). Additionally, the percentages of different COG categories (assigned to each gene by using EggNOG mapper) were plotted among the different fractions (Figures 2C,D), and the Log2 of the ratios between the Core and the "Accessory" (the combination of "persistent," shell, cloud, and unique gene fractions) among the different genomes were represented into boxplots (Figures 2E,F). The pangenome accumulation curves, fitted to the Power Law equation (Figures 2C,D), suggest that both pangenomes are open ($\gamma > 0$, see Table 1), a feature that may reflect the contact of *P. gingivalis* and *P. gulae* with several other microorganisms in a complex environment (Rouli et al., 2015), as the mammalian oral cavity. Comparing the different functional COG categories suggests that categories associated with biosynthesis and metabolism (represented by letters C to H) are more represented in the Core and persistent fractions in both *P. gingivalis* and *P. gulae* pangenomes. Moreover, the calculation of the Log2 ratios between the Core and "Accessory" fractions for both pangenomes (Figures 2E,F) showed that categories C, G, H, I, J, and M (representing the COG categories "Energy production and conversion," "Carbohydrate transport and metabolism," "Coenzyme transport and metabolism," "Lipid transport and metabolism," "Translation, ribosomal structure and biogenesis," and "Cell wall/membrane/envelope biogenesis," respectively) have a median Log2 (Core/Accessory) higher than 1. In the case of the *P. gingivalis* pangenome, categories P ("Inorganic ion transport and metabolism"), O ("Posttranslational modification, protein turnover, chaperones"), and E ("Amino acid transport and metabolism") have also higher Log2(Core/Accessory) ratios. COG categories with higher Log2 (Core/Accessory) ratios in *P. gulae* only were D ("Cell cycle control, cell division, chromosome partitioning"), F ("Nucleotide transport and metabolism"), and Q ("Secondary metabolites biosynthesis, transport and catabolism"). In all those cases, the comparison between the Core and Accessory groups was significant ($p < 0.01$, two-sided Wilcoxon Rank Sum test in R). These results suggest that, in general, housekeeping and central metabolism/ transport functions were more represented in the Core than in the Accessory gene set. This feature has been observed in other studies with other bacterial models (Gaba et al., 2020; Hyun et al., 2022). In contrast, genes without any COG hit (marked as "@" category in Figure 2) were most prevalent in the accessory fraction of the pangenome in both *P. gingivalis* and *P. gulae*. In the case of *P. gulae*, category X ("Mobilome: prophages, transposons") was also more represented in the accessory than in the core gene fraction of the pangenome. In *P. gingivalis*, genes from category V ("Defense mechanisms") had more negative Log2(Core/Accessory) ratios, whereas category X was not suitable for Log2 calculation, since the Core fraction had zero genes in that category (Figure 2C). It is worth noticing that differences from data from categories X, V, and without COG between Core and Accessory groups were significant as well ($p < 0.01$, two-sided Wilcoxon Rank Sum test).

The core gene fractions from *P. gingivalis* and *P. gulae* contained common and differentiative features. In order to compare them to define which functions were exclusive to the core fractions in each organism, a bidirectional comparison was performed with representatives from each gene from the core between those two species (Figure 3). This comparison showed a set of 1,199 bidirectional hits, and a total of 136 and 168 genes without a clear bidirectional hit in the core pangenome components of *P. gingivalis* and *P. gulae*. Respectively. Among the functional categories observed in the common and differential core contents, two cases are worth noticing between *P. gingivalis* and *P. gulae*. First, in the first one, there are more genes associated with category V ("Defense mechanisms") in *P. gulae* than in *P. gingivalis*; some of those functions are associated with a Cas-Cmr CRISPR system, several genes for multidrug efflux pump AcrA/B proteins and ABC-type antimicrobial peptide transport system, permease component (SalY)- encoding genes. In counterpart, categories H and O ("Coenzyme transport and metabolism" and "Posttranslational modification, protein turnover, chaperones," respectively) are more present in the core of *P. gingivalis* than in *P. gulae*; some of those genes encode for ABC systems for siderophores, a set of serine proteases and a cation-transporting P-type ATPase (data not shown).

10.3389/fmicb.2023.1226166

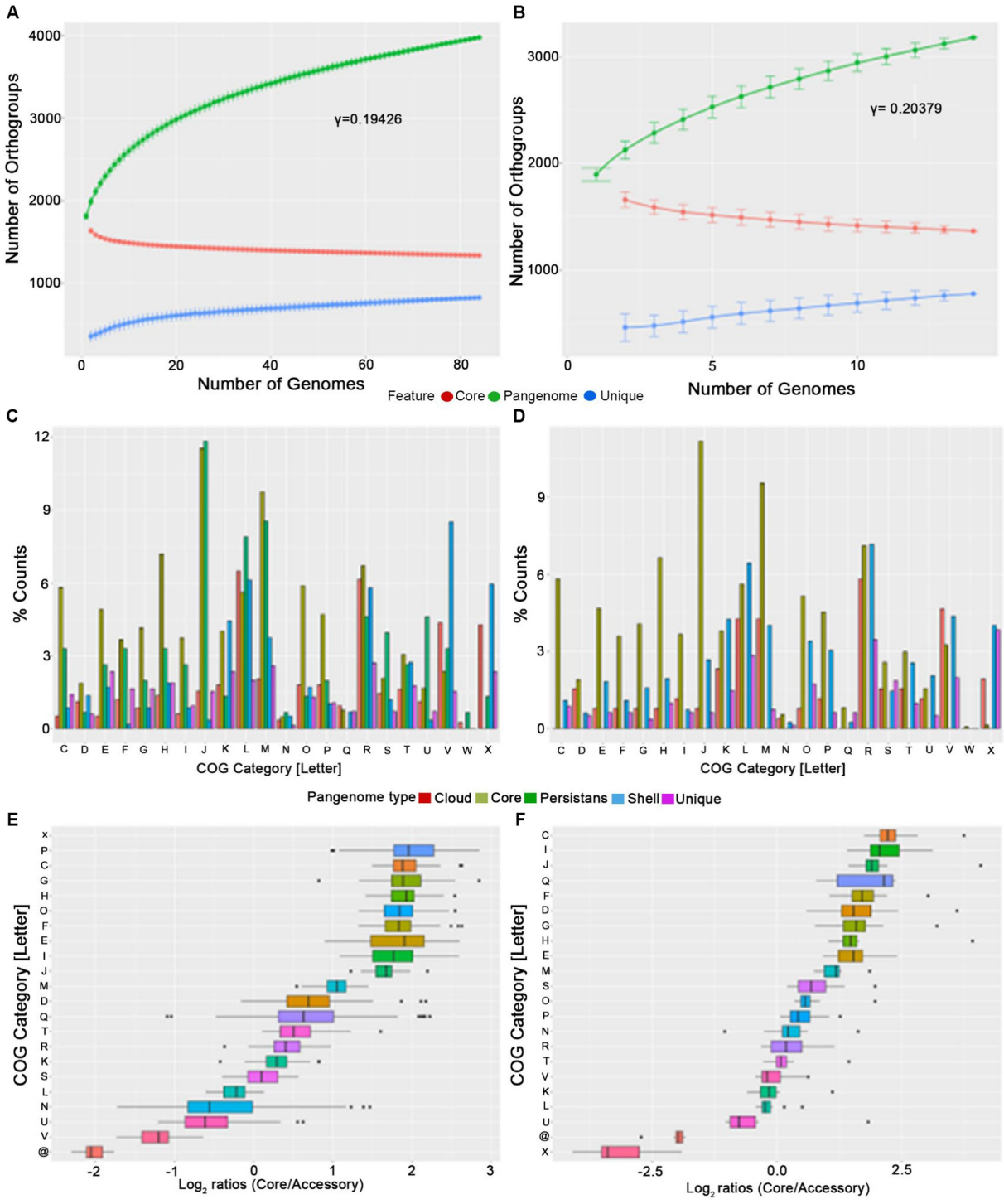


FIGURE 2
Pangenome analysis of *P. gingivalis* **(A,C,E)** and *P. gulae* **(B,D,F)**. The analysis included pangenome accumulation curves **(A,B)**, a barplot of the relative percentage of counts for each COG category **(C,D)**, and the Log2 ratios of the percentage of core vs. accessory genes for each category **(E,F)**.

3.3. Gene gain/loss model in the *Porphyromonas gingivalis*/*Porphyromonas gulae* ancestry

Our phylogenomic analysis of the *Porphyromonas* genus confirmed that *P. gingivalis* and *P. gulae* are strongly related, as well as that *P. loveana* is also directly related to them (Figure 1). In order to check about the genetic gain and loss events during divergence between *P. gingivalis* and *P. gulae*, the orthogroup and phylogenetic information from 84 *P. gingivalis* and 14 *P. gulae* genomes (including one *P. loveana* genome as an outgroup) were used to create a gene gain-loss model to represent the number of genes present in the

10.3389/fmicb.2023.1226166

predicted common ancestors across the phylogeny, by using the tool Count with Wagner parsimony analysis. The gain-loss model tree for this set of genomes is shown in Figure 4A (detailed version in Supplementary Figure S2), where each node has the number of genes analyzed, as well as the number of genes gained and lost. A set of 1,662 gene families is predicted in the ancestor of the divergence between *P. gulae* and *P. gingivalis* lineages; in this context, the last common ancestor (LCA) of all 84 *P. gingivalis* genomes contained 1,661 genes, involving the loss of 95 genes and the acquisition of 94 genes in comparison with the previous ancestor. In the LCA of *P. gulae*, 1,633 genes were found in the predicted ancestor, with a gain of 35 genes and a loss of 65 genes. Across the intraspecific branches in the *P. gulae*/*P. gingivalis* tree, relative gene gain/loss events were relatively low, occurring with a higher frequency of gain and loss events when nodes are closer to the terminal leaves (Supplementary Figure S2).

TABLE 1 Main statistics from pangenome analysis for *P. gingivalis* and *P. gulae*.

	P. gingivalis	*P. gulae*
Total genomes	84	14
Total Pangenome Genes	3,976	3,179
Core genes (in 100% of genomes)	1,335	1,367
Soft-core genes (<100%, more than 95%)	145	0
Shell genes (<95%, more than 15%)	550	786
Cloud genes (<15%, at least in two genomes)	1,124	247
Unique genes (in just one genome)	822	779
gamma value (power law equation)	0.19426	0.20379

In order to understand which classes of functions were acquired in the differentiation process between *P. gingivalis* and *P. gulae*, the functions of the genes found in the LCAs of all *P. gingivalis* and/or *P. gulae* were annotated with EggNOG mapper using COGs (Clusters of Orthologous Genes) categories at the *root* level. Considering the predicted contents of the LCAs for *P. gingivalis* and *P. gulae*, the next step is to define the content acquired and missed during the formation of those predicted ancestors. According to COGs assignments, the gained genes in the LCA for *P. gulae* were mainly associated with "K" and "L" categories (Figures 4B,C), related to "Transcription" and "Replication, recombination, and repair," with 0.18 and 0.18%, respectively. However, this LCA also suffered the loss of several genes involved in the "L" category (0.61%), and in other more generic categories such as "S" ("Function unknown"), as well as genes without any category. Some categories, such as "E" (Amino acid transport and metabolism), were more acquired in this predicted ancestor. Some of the categories with only loss were "O" and "P" (Posttranslational modification, protein turnover, chaperones, and Inorganic ion transport and metabolism), with one and four genes.

All the 94 acquired genes in the LCA of *P. gingivalis* belonged to the Unknown function category (Figures 4D,E); on the other hand, the missed genes with no COG category in this node corresponded to 20 OGs. As well as *P. gulae* "K" and "L" were the won categories with high values (0.42 and 0.60%). Interestingly, category "E" (related to amino acids) has more OGs loss compared to those won it and with *P. gulae*. Some categories with only won are the "D" (Cell cycle control, cell division, chromosome partitioning) and "J" (Translation, ribosomal structure, and biogenesis). As the *P. gulae* LCA, categories "O" and "P," have been lost (0.12 and 0.36%).

A BlastP analysis against SwissProt database for the genes acquired by *P. gulae* and *P. gingivalis* LCAs was performed to obtain more clues about the function of those genes. Results showed that 34 out of 36 proteins for *P. gulae* and 85 out of 94 for *P. gingivalis*, were predicted

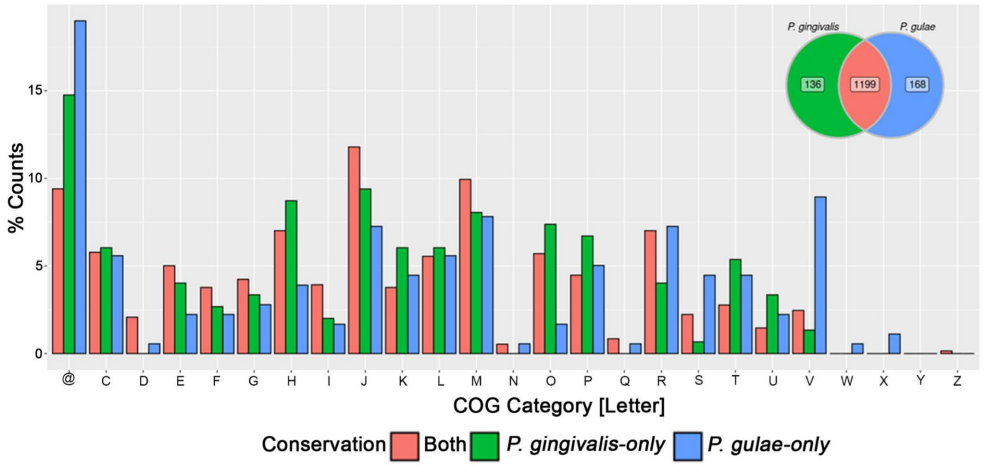


FIGURE 3
Comparison of the pangenome core of *P. gingivalis* and *P. gulae*. Bidirectional hits between the pangenome cores of each organism were represented in the Venn diagram (upper right), as well as the gene representatives without bidirectional hits in either *P. gingivalis* or *P. gulae*. The barplot of the relative percentage of counts for each COG category represented the genes found in the core from both organisms or found only in each pangenome core. @: Genes without any COG.

10.3389/fmicb.2023.1226166

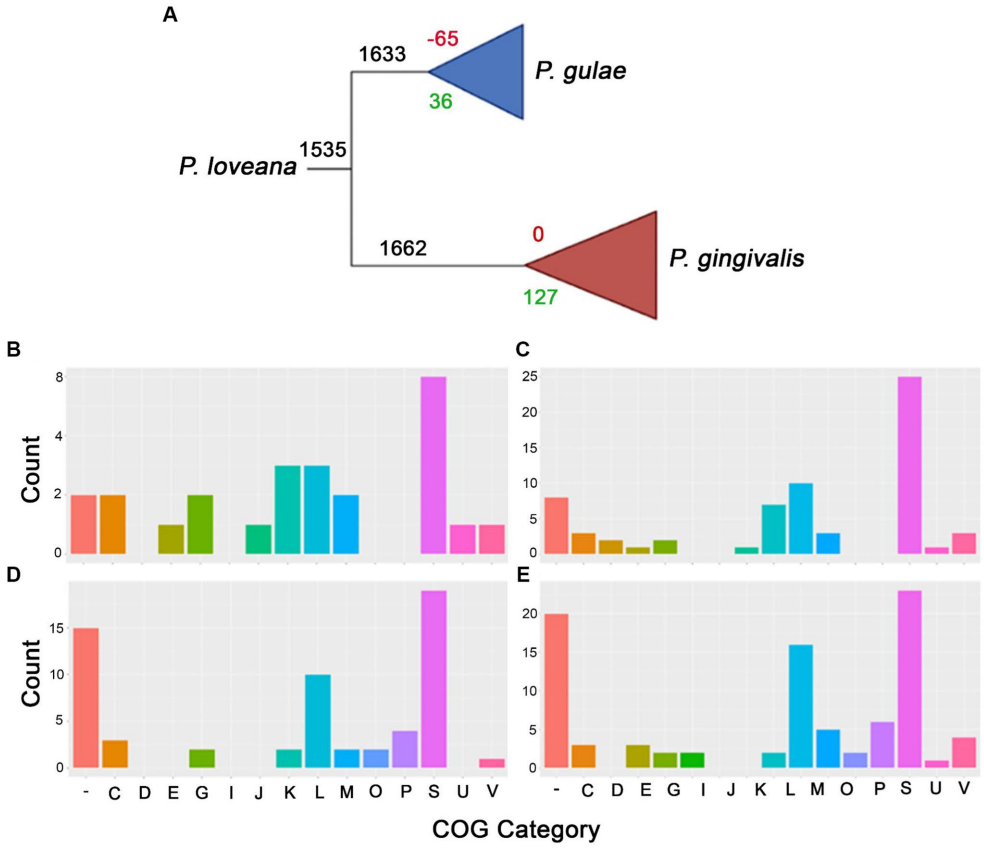


FIGURE 4
Gain and loss model genes representation for *P. gingivalis* and *P. gulae* comparison. **(A)** Collapsed phylogenetic tree representing the total amount of genes for each branch (black), as well as the gain (green) and loss (red) genes. The bottom panel represents the COGs assignment for each branch in the collapsed tree, i.e., gained and missed genes for *P. gulae* ancestor [**(B,C)**, respectively], and the set of gained and missed genes for *P. gingivalis* ancestor [**(D,E)**, respectively].

as hypothetical proteins. Global inspections of this protein and their functions for *P. gulae* (data not shown) show that 19 of those 34 sequences were hypothetical proteins; the rest corresponded to genes associated with secretion systems, transcriptional regulators, post-translational modifications, enzymes involved in metabolism, unknown protein domains, mobilome proteins, and DNA interaction proteins. For *P. gingivalis*, 46 of those 85 genes were hypothetical proteins; the rest encoded genes associated with secretion systems, transcriptional regulation, post-translational modifications, stress response and pathogenicity, DNA plasmid partition, and phage interaction.

3.4. Evolutive pressure in the shared core orthogroups from *Porphyromonas gingivalis* and *Porphyromonas gulae*

The presence of shared genes in the *P. gulae* and *P. gingivalis* lineages, raises the question about the potential effect of natural selection in the different species. In order to evaluate this possibility, the shared orthogroups of *P. gingivalis* and *P. gulae* were calculated (via Orthofinder) using the 99 genomes from these organisms (and the *P. loveana* genomes as the outgroup) and the common single-copy core orthogroups were utilized for the calculation of the evolutionary selection parameters (dN, dS, t, and dN/dS ratio).

The distribution of the dN/dS ratio between *P. gingivalis* and *P. gulae* showed an expected behavior of a long proportion of genes with dN/dS < 1 (Figures 5A,B), which is a signal of purifying selection in most of the genes. The mean for the dN/dS ratio was 0.028, compared to the mean of 15.185 of dS, showing a higher rate of synonyms mutations in the OGs set (data not shown for dS alone version and Figure 5). The means of the ratios between *P. gingivalis* and *P. gulae* are similar (0.0282 and 0.0273, respectively, with no significative difference according to the Kolmogorov–Smirnov test), however, in the same way, the mean of the dS between the two species are similar (15 and 16.1, respectively), this result shows that most of the orthogroups between both species do not have strong evidence of positive selection. If the comparisons are divided by the functional category of their genes, COGs categories with the highest dN/dS ratio medians can be found (Figure 5C); the analysis of the pairwise distributions showed that the N and S COG categories ("Cell motility" and "Unknown function," respectively) harbored the highest median for comparisons. By contrast, K and Z categories ("Transcription" and "Cytoskeleton," respectively) possessed the lower medians for dN/dS values.

Considering this feature, we perform an additional analysis where each orthogroup is compared between their values for the

10.3389/fmicb.2023.1226166

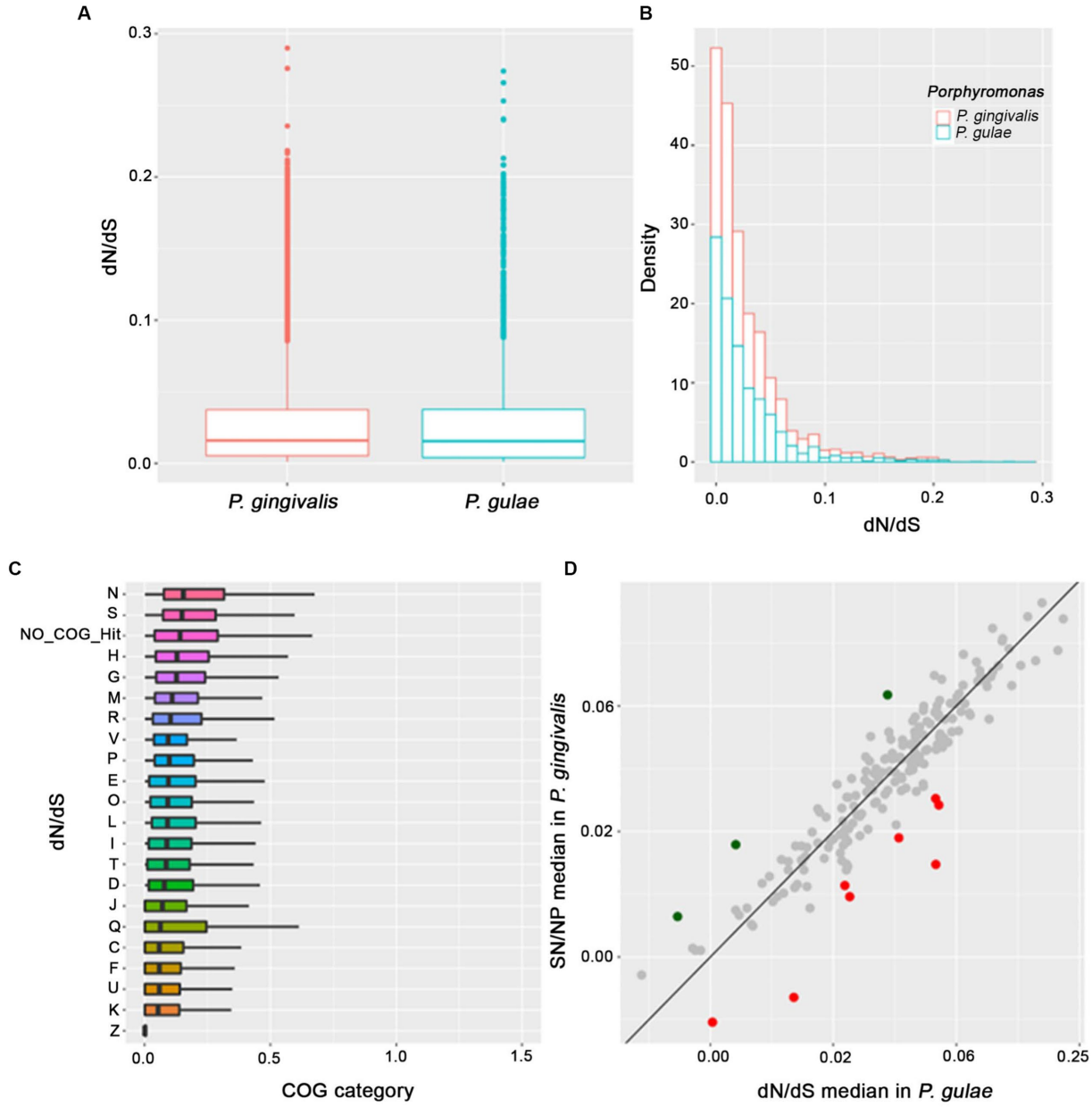


FIGURE 5
Analysis of proteome content evaluated under natural selection and the COGs category assignments. The main focus was put on the *P. gingivalis* and *P. gulae* comparison, for each omega ratio value **(A)**, the distribution of the omega value **(B)**, the distribution of each COGs and their respective values of omega ratio (ordered by the median) for each protein in the proteome **(C)** and the comparison of the omega value median of each OG (shown in gray for those without positive selection, as well as green and red for those selected as positive in *P. gingivalis* and *P. gulae,* respectively) in the comparison *P. gingivalis* versus *P. gulae* to detect signals of positive natural selection **(D)**.

P. gingivalis - *P. loveana* and *P. gulae* - *P. loveana* pairwise calculations. In Figure 5C, the scatterplot represents the median of each orthogroup considering their values for the *P. gingivalis* - *P. loveana* and *P. gulae* - *P. loveana* calculations. This comparison found 11 orthogroups with differentially higher ratios (3 for *P. gingivalis* and 9 for *P. gulae*). Using a Blastp analysis versus the SwissProt database, a more detailed description of each orthogroup was obtained (Table 2). Those genes harbored different functions, including a variety of housekeeping functions. For example, in *P. gingivalis* those genes were related to anaerobic respiration, glycogen storage, and DNA damage recognition, as 4-hydroxybutyryl-CoA dehydratase/vinyl acetyl-CoA-Delta-isomerase, glycogen synthase and Uvr ABC system protein B, respectively. On the other hand, *P. gulae* has protein related to cell wall maintenance, ion membrane translocation, anaerobic respiration, signal recognition receptor, as well as amino acid, and fatty acid metabolism.

3.5. Tajima values for *Porphyromonas gingivalis* and *Porphyromonas gulae* core orthogroups

Tajima D value is a statistical test that uses DNA data to detect natural selection (Korneliussen et al., 2013). In order to detect general

10.3389/fmicb.2023.1226166

TABLE 2 List of orthogroups found under positive selection with their respective protein function association in the comparison between *P. gingivalis* and *P. gulae*.

Species feature	OG	Function
Higher in *P. gingivalis*	OG0000556	4-hydroxybutyryl-CoA dehydratase/vinylacetyl-CoA-Delta-isomerase
Higher in *P. gulae*	OG0000576	D-alanine--D-alanine ligase
Higher in *P. gulae*	OG0000350	Ion-translocating oxidoreductase complex subunit G
Higher in *P. gulae*	OG0000351	Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit E
Higher in *P. gingivalis*	OG0000702	Glycogen synthase
Higher in *P. gulae*	OG0000267	Signal recognition particle receptor
Higher in *P. gulae*	OG0000713	Acetoacetyl-CoA:acetate/ butyrate CoA transferase alpha subunit
Higher in *P. gingivalis*	OG0000394	UvrABC system protein B
Higher in *P. gulae*	OG0000720	Acyl-CoA dehydrogenase, short-chain specific

trends in the conserved CDS in both *P. gingivalis* or *P. gulae*, the set of single-copy core genes obtained from the previously analyzed *P. gingivalis* and *P. gulae* pangenomes separately, we performed the test with 1,335 core families of *P. gingivalis* and 1,367 core families from *P. gulae*. Figure 6A shows the histogram curves densities of D value for two core orthogroups. *P. gingivalis* has a peak between 0 and −1.5. The negative values are associated with an excess of rare variation, which is interpreted as population growth or positive selection (Carlson et al., 2005). *P. gulae* Tajima's value distribution is more shifted to the negative but closer to 0 than *P. gingivalis*. Overall, *P. gingivalis* D-value distribution has a higher negative shift than genes from *P. gulae* and it was significant (p-value $<2\times10^{-16}$, Welch Two Sample t-test; Figure 6B). *P. gingivalis* may exhibit lower values either because of positive selection (more rare variants) or because a population is recovering or expanding after a bottleneck effect. The bottleneck effect refers to a reduction in the population size, resulting in the loss of genetic diversity of the species (Peery et al., 2012). In contrast, the more positive distribution of *P. gulae* values may reflect a possible population contraction or allele frequency stabilization associated with the emergence of this organism.

To observe these results in more detail, the top ten OGs with the highest and lowest Tajima values of *P. gingivalis* were retrieved (Table 3). The most positive Tajima value was 4,222; which corresponds to a gene that encodes a putative RagB, which is part of an outer membrane transporter protein that functions as a peptide importer (Madej et al., 2020). Other genes found encoded proteins with functions such as TonB-dependent receptor P3, putative tyrosine-protein kinase in *cps* locus, dipeptide and tripeptide permease B, UDP-glucose 6-dehydrogenase TuaD, Peptide chain release factor 2, Major fibrium tip subunit FimD, and polysaccharide biosynthesis/ export protein, in addition with some hypothetical proteins. These sets of genes have the most neutral/balanced selection in *P. gingivalis*. They are involved in multiple cellular processes, like heme metabolism homeostasis, heme, iron, and amino acid uptake from the host, regulation, and transport of capsular polysaccharides, protein synthesis, and function of bacterial fimbriae.

On the other hand, the top ten orthogroups with the lowest Tajima's D value (Table 4), are genes with an exacerbated proportion of rare alleles (values between −3.22 and − 3.095). The most negative one encodes a hypothetical protein, that matches (by blast analysis in the NCBI platform) with another TonB-dependent receptor, which is involved in iron and heme uptake from the host. The other proteins are: calcium-transporting ATPase 1, DUF4248 domain-containing protein, cAMP-activated global transcriptional regulator CRP, type VI-B CRISPR-associated RNA-guided ribonuclease Cas13b, 10 kDA chaperonin, transcriptional regulatory protein QseB, Acyl-CoA dehydrogenase C-terminal domain-containing protein, peptidoglycan-N-acetylglucosamine deacetylase, and GNAT family N-acetyltransferase. Their functions are related overall to iron and heme uptake from the host, calcium, and cAMP intracellular levels homeostasis, a prokaryotic defense mechanism against foreign genetic elements, regulation of biofilm formation, fatty acid metabolism, and cell wall modifications.

3.6. Virulence factor conservation and evolution among *Porphyromonas* genomes

P. gingivalis is known to produce a set of virulence factors involved in the formation of the periodontal disease (Aleksijević et al., 2022; Nuñez-Belmar et al., 2022). Among them, FimA and Mfa1 are filamentous structures essential for the interaction of *P. gingivalis* with other bacteria of the biofilm and with host cells (Nagano et al., 2018). *P. gulae* also was shown to exhibit these adhesive molecular systems (O'Flynn et al., 2015). The major and minor fimbriae exhibit adhesion properties that are suggested to play a central role in periodontitis development (Mendez et al., 2019). Other virulence factors are the gingipains and the Hemagglutinins (Hag). Hag proteins are bacterial surface glycoproteins associated with bacterial adhesion to erythrocytes and host cells (Mendez et al., 2019). For *P. gingivalis* metabolism the heme acquisition is essential, thus, hemagglutinins adhesive domain function is related to this process, facilitating the acquisition of heme. Hag-encoding genes (*hagA*, *hagB*, and *hagC*) are described as relevant virulence factors of *P. gingivalis* (Nakayama and Ohara, 2017). Gingipains are lysine- and arginine-specific proteinases secreted by *P. gingivalis*, and they are considered as one of the most important virulence factors in this organism; their principal function is the degradation of host proteins and the processing of the fimbriae subunits (Lunar Silva and Cascales, 2021). Finally, peptidyl-arginine deiminase (PDA) are enzymes previously reported in *P. gingivalis* capable of producing protein citrullination associated with their capabilities in microbe-host interaction and a role in some diseases such as arthritis (Maresz et al., 2013). The differential degree of presence/absence of orthogroups among Porphyromonads may also occur in the known virulence factors associated with *P. gingivalis* in periodontitis and other diseases. To check for other hits for a variety of the main virulence factors, a *hmmer* search for different Pfam domains (Supplementary Table S3) was performed, searching for

10.3389/fmicb.2023.1226166

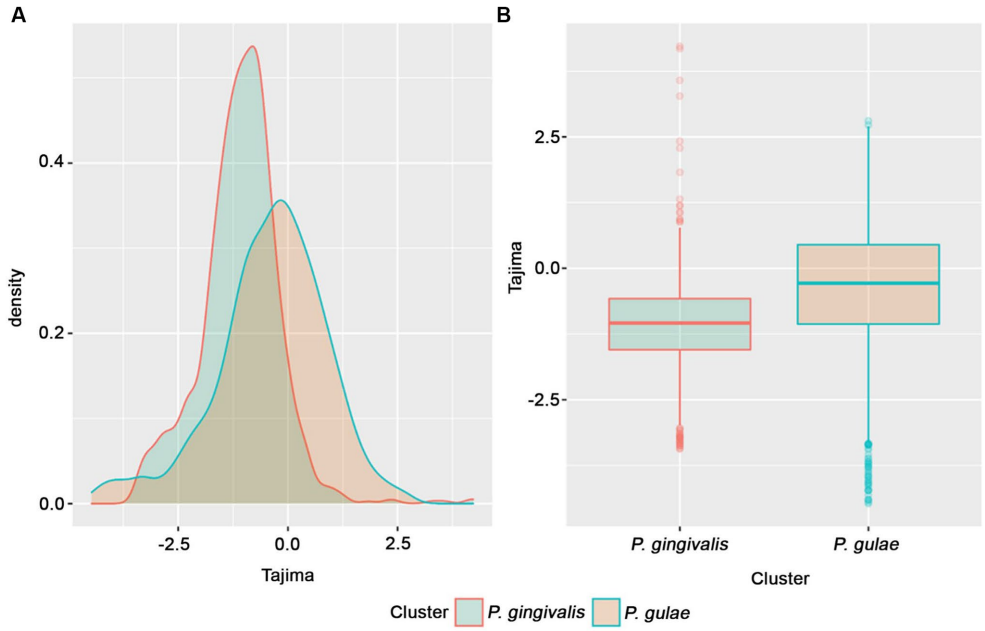


FIGURE 6
Tajima's D value distributions for *P. gingivalis* and *P. gulae* core gene families. **(A)** histogram showing density distributions for Tajima values data for *P. gingivalis* and *P. gulae*; **(B)** boxplots for Tajima values clustered by species.

families associated with previously reported virulence factors, with a posterior phylogenetic analysis of those hits. From the PFAM database, this analysis utilized the families identified as PF06321, containing sequences for FimA (genotypes I, II, and IV) and FimC; PF15495 containing Mfa1 and other related proteins; and PF010365, containing hemagglutinin and other proteins such as Lys-gingipains and Arg-gingipains.

These analyses suggested that the number of genes containing several of those domains exhibited a differentiative pattern in the group of *P. gingivalis - P gulae - P. loveana*. For example, there are more genes containing the PF06321 domain ("*P_gingi_FimA*") in those organisms than in the rest of the Porphyromonads. Hits for PF06321 and PF15495 ("*Major fimbrial subunit protein type IV, Fimbrillin, C-terminal*") domains were widely distributed among *Porphyomonas* genomes, being detected in multiple copies in *P. gingivalis*, *P. gulae*, *P. loveana,* and *P. macacae*. A phylogenetic tree base on aminoacid sequences of FimA (Figure 7) and Mfa1 (Figure 8) from the selected *Porphyoromonas* genomes. The original PF06321 FimA phylogenetic tree (Supplementary Figure S3) showed two main branches: one of them represented the FimC (proteins from *P. gingivalis, P, gulae, P. loveana, P. levii*, and *P. crevioricaris*), and the second clade represented FimA from several Porphyromonads. The FimA clade is colored by each genotype (type I, type II, and type IV in Figure 7), and represented proteins from *P. cangingivalis, P. canoris, P. levii, P uenonis, P. endodontalis, P. macacae, P. levii, P. gingivalis, P. gulae,* and *P. loveana*. Genotype II FimA proteins were detected in the trifecta *P. gingivalis, P. gulae, and P. loveana*, whereas genotypes I and IV were found in *P. gingivalis* and *P. gulae*. Interestingly, FimA homologs were also detected in other members of the genus, such as *P. endodontalis, P. levii, P. cangingivalis, P. ueonis* and *P. canoris*; moreover, the most related to the ancestral branches were found in *P cangingivalis* and *P. canoris*, suggesting that this branch could predate the origin of the *P. gulae - P. gingivalis* duo.

The minor fimbriae Mfa1 was also present in a variety of species of the Porphyromonas genus (Figure 8), being detected in *P. gingivalis*, *P. gulae*, *P. loveana*, *P. macacae*, *P. levi*, *P. uenonis*, *P. crevioricanis*, *P. gingivicanis*, *P. circumdentaria*, *P. endodontalis*, *P. somerae*, and *P. asaccharolytica*. The original tree PF15495 phylogenetic tree (Supplementary Figure S4) showed two main branches, one of them clearly composed by Mfa1 homologs. In this subtree, two linages of *P. gingivalis - P. gulae - P. loveana* were detected: One is associated with *P. macacae* and *P. levi*, which each are isolated from dog oral cavity and cow skin lesions. The other lineage is associated with *P. endodontalis*, which is isolated from human subgingival biofilm and saliva, besides has been closely associated with periapical lesions of patients with periodontitis (Lombardo Bedran et al., 2012). This phylogenetic configuration suggests that, in *P. gulae* and *P. gingivalis*, the mfa1 genes could have two origins, dictated by their transfer from other members of the *Porphyromonas* genus.

The phylogenetic tree with protein sequences filtered by PFAM PF010365 from the *Porphyromonas* dataset (Figure 9) included proteins such as the hemagglutinins and the gingipains [divided into lysine-gingipains, arginine-gingipains, proteases with an adhesive hemagglutinin-like domain (Mysak et al., 2014)], as well as a set of hypothetical proteins. Lysine-gingipains are exclusively associated with *P. gingivalis*, whereas, the Arginine-gingipains were associated with *P. loveana* and *P. gingivalis*. The hemagglutinin A lineage was represented by two branches: one with *P. gulae* and *P. loveana* proteins, and the other with *P. gingivalis* and *P. gulae* proteins. The presence of hemagglutinins and gingipains in *P. loveana* may suggest that this species could have a similar pathogenic phenotype as observed in *P. gingivalis*.

TABLE 3 List of orthogroups covering the 10 highest Tajima values of in the *P. gingivalis* core, including their protein names and functions ($p < 0.05$).

Top position	OG	Tajima value	Protein	Function
1	OG0000294	4.2226932	Hypothetical protein (RagB)	RagB is an outer membrane transporter protein that functions as a peptide importer in *P. gingivalis*.
2	OG0000293	4.17820696	TonB-dependent receptor P3	Protein involved in the acquisition of essential nutrients, such as iron and heme, from the host
3	OG0000559	3.574256432	Hypothetical protein	Unknown
4	OG0001241	3.348000822	Hypothetical protein	Unknown
5	OG0000454	3.275760793	Putative tyrosine-protein kinase in cps region	The cps (capsular polysaccharide synthesis) region is responsible for the biosynthesis of the capsular polysaccharides
6	OG0001351	2.715270824	Dipeptide and tripeptide permease B	Protein involved in the uptake of peptides that are used by the bacterium for the synthesis of various virulence factors, such as gingipains
7	OG0000872	2.418843779	UDP-glucose 6-dehydrogenase TuaD	Biosynthesis of the bacterial cell surface polysaccharide
8	OG0000873	2.287357949	Peptide chain release factor 2	Termination of protein synthesis during bacterial translation
9	OG0001381	1.90044368	Major fimbrium tip subunit FimD	Assembly and function of the bacterial fimbriae
10	OG0000455	1.826818797	Hypothetical protein (polysaccharide biosynthesis/export protein)	Assembly of the polysaccharide chain into a repeating unit, as well as its transport across the cell membrane

For the case of the PAD proteins, a phylogenetic tree using those sequences showed that *Porphyromonas* PF04371 hits were distributed in two lineages (Supplementary Figure S5). One comprises the PAD proteins from *P. gingivalis*, *P. gulae,* and *P. loveana.* The other lineage contains the putative agmatine deiminases from nearly all members of the genus. This two-branch divergence suggests that the previous protein citrullination activity involved in pathogenic features in *P. gingivalis* could also be present in *P. gulae* and *P. loveana.*

4. Discussion

4.1. The uncovered diversity of the *Porphyromonas* genus and its relationship with the *Porphyromonas gingivalis* - *Porphyromonas gulae* duo

In order to investigate how *P. gingivalis* could become a pathogen in the context of the *Porphyromonas* genus, the first aspect we needed to analyze is to validate its position in relation with its relatives. We performed with a set of 233 high-quality *Porphyromonas* genomes a combination of a phylogenomic analysis with the use of genomic metrics for the establishment of genomic species clusters (Maturana and Cárdenas, 2021). The phylogenetic structure of the genus (Figure 1) showed that there are some clades mostly composed of undescribed species, most of them already classified by the GTDB (Supplementary Table S1). This feature has become common in several genera among bacterial and archaeal species as a reflection of their novelty in specialized databases (Chaumeil et al., 2022). Additionally, the phylogenomic analysis may propose new species or genera from previously described and undescribed strains (Sawana et al., 2014; Gupta et al., 2020). In this case, several undescribed *Porphyromonas* species could be representatives of different oral cavity inhabitants, as well as colonizers from other body sites, with their roles in host health and disease yet to be understood. For example, two undescribed species (mentioned as "oral taxon 278" and "oral taxon 275") detected in plaque samples (Mei et al., 2020), and the misnamed strain 60–3 (named as a *P. uenonis* strain, despite it is not part of the species) may represent some relevant new species remaining to be officially named and described.

The phylogenomic analysis of the *Porphyromonas* genus confirmed the strong relatedness with *P. gulae* [formerly a *P. gingivalis* subspecies associated with non-human hosts, see (Fournier et al., 2001)], but also showed a strong relatedness with *P. loveana*, a strain isolated from the oral cavity of a musky rat kangaroo (Bird et al., 2016). The exact role of this latter species in oral health and disease in its host remains unclear. However, as seen in later results (see Section 3.6), several virulence factors were also conserved in this species, as well as the *P. gingivalis* - *P. gulae* duo. This may suggest a potential role in the periodontal disease of *P. loveana* in their marsupial host. This may also suggest two potential scenarios: (A) the divergence between *P. gingivalis*/*P. gulae* duo and *P. loveana* are the potential origins of the periodontal disease and it is as old as the divergence of early mammals and marsupials [estimated to occur between 130 and 170 million years

10.3389/fmicb.2023.1226166

TABLE 4 List of orthogroups covering the 10 lowest Tajima values of in the *P. gingivalis* core, including their protein names and functions ($p < 0.05$).

Top position	OG	Tajima value	Protein	Function
1	OG0001008	−3.223375698	Hypoyhetical protein (TonB-dependent receptor)	Protein involved in the acquisition of essential nutrients, such as iron and heme, from the host
2	OG0000740	−3.211357054	Calcium-transporting ATPase 1	Ion pump involved in calcium homeostasis and tolerance to high concentrations of calcium
3	OG0000475	−3.206282293	Hypothetical protein (DUF4248 domain-containing protein)	Unknown, it is suggest that is involved in protein–protein interactions, DNA binding, or enzymatic activities
4	OG0001040	−3.204872637	cAMP-activated global transcriptional regulator CRP	Regulation of respond to chages in intracellular levels of the signaling molecule cyclic AMP (cAMP)
5	OG0000392	−3.186069821	Hypothetical protein (type VI-B CRISPR-associated RNA-guided ribonuclease Cas13b)	Prokaryotic defense mechanism against foreign genetic elements such as viruses and plasmids
6	OG0000511	−3.158209672	10 kDa chaperonin	Folding of newly synthesized proteins
7	OG0000636	−3.155269224	Transcriptional regulatory protein QseB	Response regulator that is activated by the QseC sensor kinase, involved in the regulation of biofilm formation and antibiotic resistance
8	OG0000667	−3.137577035	Hypothetical protein (Acyl-CoA dehydrogenase C-terminal domain-containing protein)	Catalyze the first step in the breakdown of fatty acids
9	OG0001389	−3.121169508	Peptidoglycan-N-acetylglucosamine deacetylase	Modification of peptidoglycan, an structural molecule of bacterial cell wall
10	OG0001157	−3.095009015	Hypothetical protein (GNAT family N-acetyltransferase)	Modification of a variety of cell wall components, including lipopolysaccharides and capsular polysaccharides

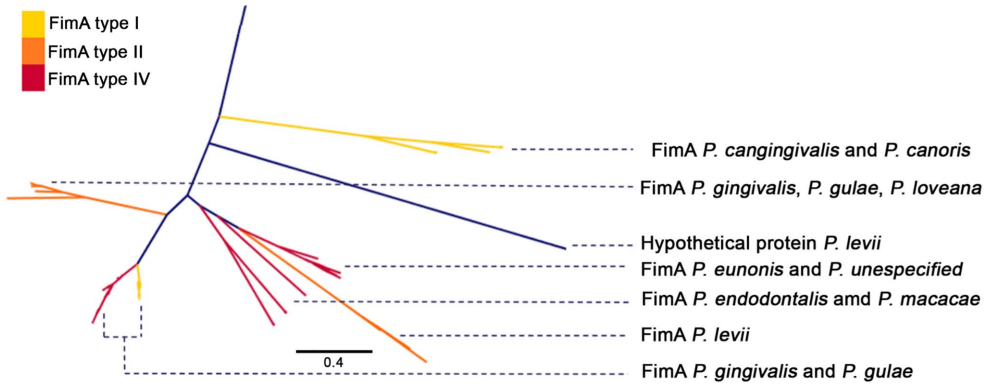


FIGURE 7
Phylogenetic tree, based on the FimA amino acid sequences (PFAM PF06321) of the *Porphyromonas* genus. FimA branches are colored according to FimA genotype (type I yellow, type II orange, and type IV red), and labeled with *Porphyromonas* species.

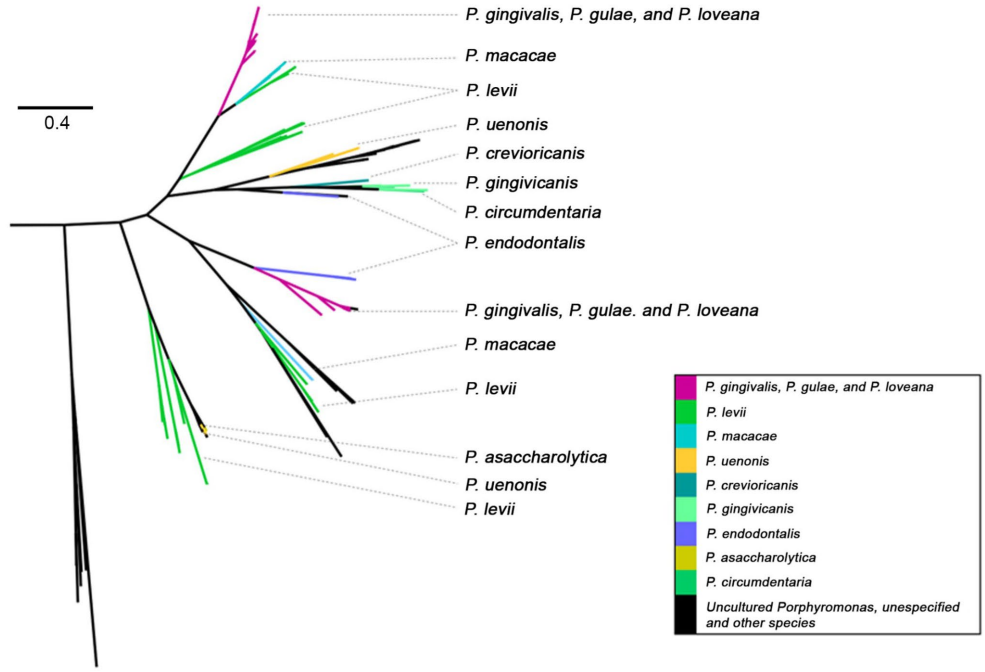


FIGURE 8
Phylogenetic tree, based on Mfa1 amino acid sequences (PFAM PF15495) of *Porphyromonas* genus. Tree branches were colored according to their taxonomic origin among *Porphyoromonas* species.

ago (Cifelli and Davis, 2003)]; or (B) periodontal disease is potentially rooted in the mammals, regardless of the date of origin of the *Porphyromonas* species associated to it. To address these affirmations, it will be necessary to sequence more periodontitis-associated Porphyromonads in several other mammals.

4.2. Pangenome openness of *Porphyromonas gingivalis* and *Porphyromonas gulae* and the potential role of gained genes among LCAs during species divergence

Pangenome approaches allow us to know which genetic repertoire of a bacterial group, especially discover essential genes for the survival or even all the virulence factors in a specific bacterial species (Bosi et al., 2015; Rouli et al., 2015). Our analysis shows that the pangenome of *P. gingivalis* and *P. gulae* are in an openness stage (Figure 2; Table 1), suggesting that the gene content from both species will be accepting more new gene families while new genomes are added to the analysis. The genomic core for both organisms was enriched housekeeping functional categories whereas the accessory fraction of the pangenomes was enriched by genes without COG and associated with mobile elements and defense mechanisms. This adaptation of the accessory core in both species is a previously observed feature, reflecting a recent history of mobile element transfer and the utilization of fitness factors (Brockhurst et al., 2019). The presence of unknown genes Interestingly, some proteins present in the *P. gulae* were related to phage defense like CRISPR-Cas system associate proteins. Recent studies have suggested the presence of different phages in *P. gingivalis,* raising the potential for cross-phage infection between relative species like *P. gulae* and *P. gingivalis* (Matrishin et al., 2022).

The gene gain/loss analysis showed a comparable number of gene gain and gene loss events during the divergences between *P. gingivalis, P. gulae,* and *P. loveana*, with the number of gene gain/loss events after the speciation process. Since that previous evidence showed that selection could be a significant driver of gene loss and reductive genome evolution in bacterial species (Koskiniemi et al., 2012) and that gene content reconstruction can reflect the evolution of microbial pangenomes in a specific taxon (Iranzo et al., 2019; Maturana and Cárdenas, 2021), events observed during *P. gingivalis* - *P. gulae* bifurcation suggested a discrete role of gene gain/loss in this species delimitation, as a signal that those two species are very similar in the evolutionary pressure produced by their respective environments.

To our knowledge, there is no previous report about gain and/or loss events describing *Porphyromonas* evolution, even if it is focused only on *P. gingivalis*. The fact that several genes gained by the LCA of *P. gingivalis* were related to metabolic enzymes and transcriptional regulators, genes with these specific characteristics were categorized as essential for fitness in *P. gingivalis* (Miller et al., 2017). Moreover, one of the genes was involved in the nitro-groups (nitroreductase family protein). In the oral microenvironment, nitrosative stress could be challenging due to the high intake of dietary nitrate, some genes are necessary for the growth of this oral pathogen (Lewis et al., 2012). Some of the gained genes by both LCAs were related directly to the mobilome of this pathogen, which involved genes responsible for relaxase/mobilization nuclease, transposase, plasmid mobilization relaxosome, integrase, and transposons. All those elements play an important role in genetic rearrangement and are considered driving forces of bacterial diversification (Olsen et al., 2018). Two proteins that were identified as T9SS type A were gained in this node, T9SS

10.3389/fmicb.2023.1226166

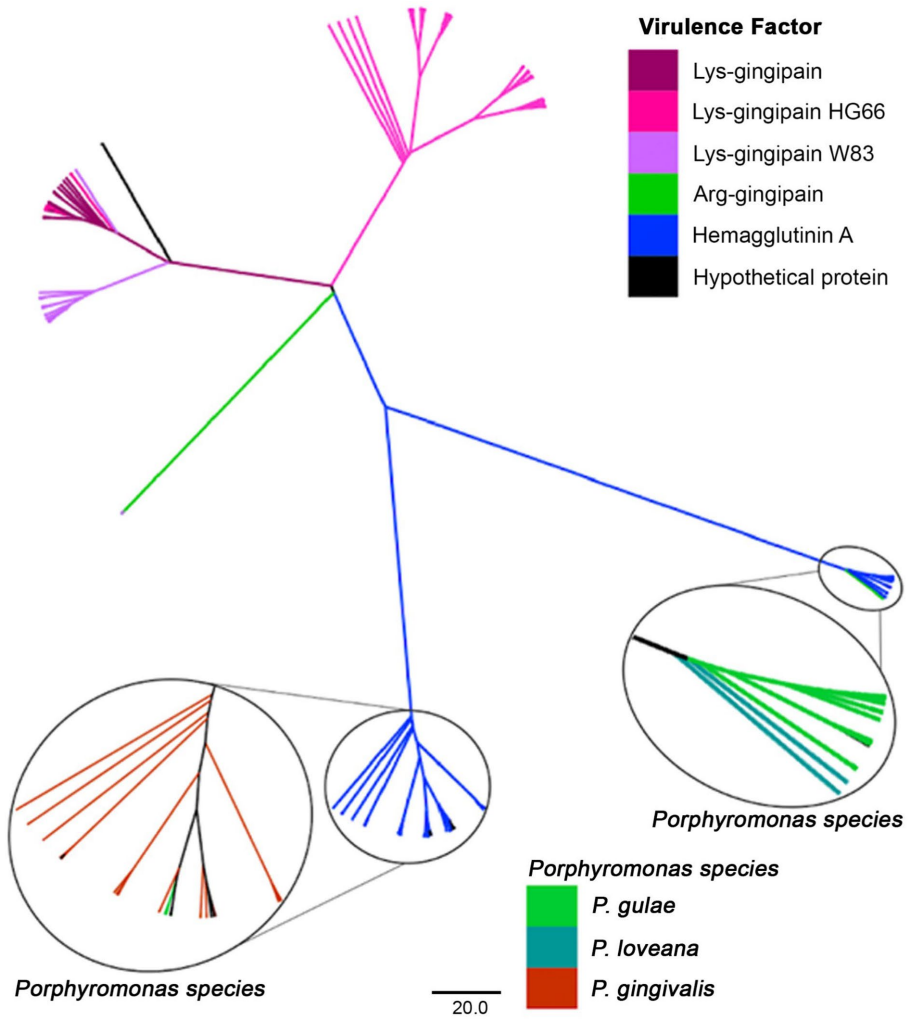


FIGURE 9
Phylogenetic tree, based on the hemagglutinin and gingipain amino acid sequences (PFAM PF010365) of *Porphyromonas* genus. Tree branches were colored according to their putative function, separating among hemaglutinins, Lys- or Arg-gingipains. Taxonomic origin is highlighted for the hemaglutinin clades.

translocates proteins to the outer membrane, which is important for virulence factors. These proteins play an important role in cell survival and fitness in response to the microenvironment (Lasica et al., 2017). One important virulence factor was reported, glycosyltransferase, which is important for capsular polysaccharide biosynthesis (Naito et al., 2008).

In the case of *P. gulae*, we found at least two type IX secretion systems which are interesting in the way that these kinds of systems are present in some bacteria species and could play two main roles, which are gliding motility or a weapon for some pathogens, as *P. gingivalis* described above. To our knowledge, this kind of system was not described previously in *P. gulae*, delving into the sequences inside the genome of some members of *Porphyromonas* could reveal some particular features. Proteins related to transcription factors process or interact with DNA as well as an energy process, where were pNresent between the genes gained by the *P. gulae* LCA. Some elements like relaxosome protein, which is a protein that allows the conjugation process, a type of horizontal transfer gene, this kind of system could play an important role in the genetic transfer of information between members of *Porphyromonas* (Watanabe et al., 2017). Finally, the LCA of *P. gulae*, like *P. gingivalis*, also gained a nitroreductase protein. These two species are members of the oral microbiome in mammals, as we described above the oral microenvironment is rich in nitro-groups.

4.3. Differential gene selection across *Porphyromonas gingivalis* and *Porphyromonas gulae* may be part of the specialization process

The analysis of dN/dS ratios is a method to study evolutionary pressure among macroevolution in gene families (Vitti et al., 2013). The measure of those values was performed for the shared core of *P. gingivalis* and *P. gulae*, considering *P. loveana* as the outgroup. The observation that the "N" and "S" categories contained the highest median for dN/dS data, could reflect a slightly higher diversification effect for genes associated with those categories; genes associated with

category N were involved in functions such as cell cycle control, cell division, and chromosome partitioning, including proteins like ParA (Bignell and Thomas, 2001), a member of a large group of P-loop ATPases with a deviant Walker A motif, involved in DNA partitioning (Mishra and Srinivasan, 2022). Proteins like this could be under the effect of positive selection due to they are essential for proper DNA partitioning and cell division, in the same way, the effect of natural selection could give advantages to ensuring accurate DNA and improve fitness (Liu et al., 2008). Another protein found in category N was part of a complex involved in the translocation of lipopolysaccharide (LPS), the permease protein Lpt (Lpt). The translocation of LPS to the outer membrane could be related to cell motility and membrane dynamics (Veronika et al., 2011). LPS is an important virulence factor for *P. gingivalis* and/or *P. gulae*, the effect of positive selection could improve the impact of this virulence factor in the oral microbiome (Dashper et al., 2017). The S category is the second with the highest median; genes from this category were generally annotated as membrane proteins and lipoproteins. Further investigation is necessary to characterize those genes.

The *P. gulae* and *P. gingivalis* lineages could have different evolutionary pressures. The comparison of their respective values for the set of each shared orthogroup showed that only a few gene families contained a considerable signal making them candidates to have a more positive selection in one lineage than in another: three belonged to *P. gingivalis* and six to *P. gulae*. The function of some of these gene families could give some clues about the specialization of each lineage. For example, genes under positive selection in *P. gingivalis* were 4-hydroxy butyryl-CoA dehydratase/vinyl acetyl-CoA-Delta-isomerase, Glycogen synthase, and UvrABC system protein B. The first protein is involved in the oxidation of fatty acids mostly in anaerobic microorganisms, this protein has a relationship with *P. gingivalis*, *Clostridium aminobutyricum,* and *Archaeglobus fulgidus* (Gerhardt et al., 2000) and its positive selection may respond to the need to improve its fitness to synthesize lipids in their environment, in comparison with *P. gulae*. Glycogen synthesis could influence the survival mechanism of microorganisms (Park, 2015); previous studies in other pathogens such as *Vibrio cholerae* showed that glycogen biosynthesis is a fitness factor under nutrient-limiting conditions, such as the aquatic environment (Kamp et al., 2013). The UvrABC system protein B is part of a complex involved in DNA repair and with a relevant role in damage resistance in *P. gingivalis* during oxidative stress (Henry et al., 2012). The differentially higher positive selection may respond to the need to improve fitness by diversifying selection in those genes when the *P. gingivalis* lineage is compared with *P. gulae*.

4.4. Tajima's D values suggest evidence of a bottleneck effect in the human host

As mentioned earlier, *P. gingivalis* and *P. gulae* are closely related phylogenetically (O'Flynn et al., 2015; Fujiwara-Takahashi et al., 2020). Additionally, the closest relative to them is *P. loveana*, a species isolated from the oral cavity of Australian marsupials (Bird et al., 2016). *P. gingivalis* and *P. gulae* emerged from a common ancestor, and although both are associated with periodontitis, they differ in host specificity. *P. gingivalis* is isolated from human samples and *P. gulae* from other mammal hosts (Fujiwara-Takahashi et al., 2020). Such different host specialties could be associated with changes in neutrality among core genes.

We performed the Tajima D statistical test with core gene families of *P. gingivalis* and *P. gulae* and generated curves and boxplots with all the Tajima's scores for gene families of those organisms (Figure 6B). The differences among the Tajima values between species were significant ($p < 2 \times 10^{-16}$). *P. gingivalis* distribution presented a distribution towards lower Tajima's D values than *P. gulae*., exhibiting less neutrality in their core orthologs (higher frequencies of rare alleles). This observation may be explained due to an expansion after a bottleneck effect.

The bottleneck effect refers to a past event that causes the reduction of the population since the outcome is the loss of genetic diversity, with impacts on the appearance of rare alleles (Peery et al., 2012). The foundation effect is caused when a small group of individuals establish a new population. This causes the emergence of a new population with less genetic diversity and an increment in the frequency of rare alleles (Nielsen, 2001). In the case of *P. gingivalis*, this may be explained due to its specialization as the only Porphyromonad causing periodontitis in the human oral cavity, exhibiting an important site, host, and activity specificity (Bostanci and Belibasakis, 2012). In contrast, its relative *P. gulae*, shows similar characteristics and a role in periodontal disease but is associated with a wider host range, as it is isolated from different mammals. Considering the pathogenicity and specificity of the hosts of both species, as well as the findings of our study, we dare to hypothesize that *P. gingivalis* experienced the bottleneck effect, followed by an amplification of rare alleles, during its colonization and establishment in the human oral cavity. We speculate that both *P. gingivalis* and *P. gulae*, from their same common ancestor, contained the virulence factors necessary for periodontal infection and after a speciation event, they evolved to adapt to their hosts, being *P. gingivalis* the one whose evolved to a different species in the human body. However, since the numbers of *P. gingivalis* and *P. gulae* genomes were unequal, sequencing of more *P. gulae* genomes and their posterior analysis can help to confirm these allele frequency differences.

4.5. Virulence factor conservation among *Porphyromonas* species shows a common pattern in lineage emergence for periodontal disease

Phylogenetic analyses for some of the most relevant virulence factors involved in periodontitis showed patterns that could be explained by the evolution of *Porphyromonas* species, as well as by their association with their hosts. For example, phylogenies for the major fimbriae (FimA) and the minor fimbriae (Mfa1) showed the co-clustering of *P. gulae, P. gingivalis* and *P. loveana* sequences, in agreement with the common evolutionary origin for those organisms. In the case of the different genotypes (I, II, and IV), the potentially closest clade to the root of the group is the type II genotype, which also includes the ortholog from *P. loveana*. The furthest branch contained FimA genotypes I and IV, encoded by *P. gingivalis,* and *P. gulae* strains. Current evidence has suggested that type I genotype is associated with samples from patients with periodontal health,

whereas genotypes II and IV are associated with samples from patients diagnosed with periodontitis (Wang et al., 2020). This may suggest that the rise of the attenuated type I could be posterior to the emergence of the pathogenic type II, aspect that could ensure the early capability of *P. gingivalis* and *P. gulae* to succeed in their role in periodontal disease in their respective niches. In contrast, the late adaptation into genotypes I and IV may respond to dynamics of host specificity adaptations as long as *P. gulae* and *P. gingivalis* are adjusting to their new niches. Previous studies suggested that genotype IV fimbriae are associated to increased adhesion efficiency in comparison with genotype I (Mendez et al., 2019); this behavior may reflect a late adaptation to increase adhesion abilities after the formation of the "soft" genotype I phenotype.

The presence of FimA orthologs in other Porphyromonads (*P. cangingivalis, P. canoris, P. levii, P uenonis, P. endodontalis, P. macacae, P. levii*) expand the previous findings from (Fujiwara-Takahashi et al., (2020), which only considered *P. gulae* and *P. gingivalis*. In fact, one early clade in the FimA phylogeny contained orthologs from *P. cangingivalis* and *P. canoris*, species isolated from canine oral cavities. *P. cangingivalis* is isolated from health and early periodontitis-associated subgingival plaque, and *P. canoris* from subgingival plaque associated to periodontitis (Collins et al., 1994; Love et al., 1994). This finding suggests that the origin of FimA may predate the adaptation of the role of Porphyromonads in periodontal disease.

A similar "trifecta-enriched" phylogenetic pattern was also observed with Mfa1, a protein expressed by *P. gingivalis* on its surface. The primary function of the fimbrial systems is to adhere to host epithelial cells and colonize subgingival tissue. However, Mfa1 is also involved in microorganism auto-aggregation, microbe-microbe interactions, and suppression of the host immune response (Arjunan et al., 2016), via the evasion of autophagy-lysosome function of myeloid dendritic cell (DC) when it internalizes *P. gingivalis* (El-Awady et al., 2015). Mfa1 is also able to induce osteoclastogenesis by interacting with the Toll-like receptor of murine macrophage (RAW264 cells) *in vitro* (Suzuki et al., 2022). Osteoclastogenesis and consequent bone resorption is the hallmark of periodontitis (Hajishengallis and Lamont, 2016). Additionally, the *in vitro* interaction between *P. gingivalis* Mfa1 and the streptococcal antigen I/II receptor (SspA/B) of the early colonizer *Streptococcus gordonii* was demonstrated (Roky et al., 2020).

A phylogenetic tree of Mfa1 proteins was constructed from the PF15495 hits found in *Porphyromonas* genomes. Congruent Mfa1 homologs were detected in *P. gingivalis, P. gulae,* and *P. loveana,* in addition to *P. macacae, P. levi, P. uenonis, P. crevioricanis, P. gingivicanis, P. circumdentaria, P. endodontalis,* and *P. asaccharolytica*. Whereas *P. macacae* and *P. cangingivalis* were identified in the oral cavity of a variety of mammals (dogs, cats, monkeys, and bovines) (O'Flynn et al., 2015; Gabarrini et al., 2018). Additionally, *P. uenonis,* and *P. asaccharolytica* have been detected in the cervicovaginal tract and appendicitis- or peritonitis-associated samples (Finegold et al., 2004; Lithgow et al., 2022) or in diabetic patients with foot infections (Imirzalioglu et al., 2014). Another interesting aspect of the Mfa1 phylogeny is that homologs from the trio *P. gingivalis* - *P. gulae* - *P. loveana* are distributed in two separated clades suggesting that in those organisms, the origin of this protein was carried out twice. The separation of those clades is a resemblance of the existence of two proposed genotypes in the Mfa1 genes observed in *P. gingivalis* and *P. gulae*, the so-called 70-kDa and 53-kDa types (Fujiwara-Takahashi et al., 2020). Structural and more detailed evolutionary studies of those separated clades are worth a more profound next study.

Organisms with detected FimA and Mfa1 homologs can be grouped (Figure 10). More species harbored Mfa1 compared to FimA. Two species, *P. cangingivalis* and *P. canoris*, contained FimA but not Mfa1, and four organisms (*P. crevioricanis, P. gingivicanis, P. circumdentaria,* and *P. asaccharolytica*) contained Mfa1 and not FimA. *P. crevioricanis* and *P. gingivicanis* were isolated from the gingival crevicular fluids of dogs (Hirasawa and Takada, 1994), whereas *P. circumdentaria* and *P. asaccharolytica* were isolated from the gingival margins of cats (Love et al., 1992), and from a variety of nonoral clinical samples (Shah and Collins, 1988). These differences in gene content may imply that those organisms evolved to contain the minor fimbria to function in other roles outside of the periodontal disease. A more profound study can help to elucidate the properties of those Mfa1 homologs.

The evolutionary patterns of the hemagglutinin-containing proteins are more straightforward. The phylogenetic compartmentalization between arginine-gingipains (RgpA), lysine-gingipains (Kgp), and hemaglutinins are clear. Interestingly, this tree only included proteins from the trifecta *P. gingivalis* - *P. gulae* - *P. loveana*. Gingipains were found only in *P. gingivalis* and in very few strains of *P. gulae*, whereas hemagglutinins were found in the three related organisms. The tree divided the Lys-gingipains into two lineages, associated with HG66 and W83 strains, respectively. The HG66 is an avirulent *P. gingivalis* strain exhibiting the type I genotype FimA, whereas the strain W83 is a more virulent variant, exhibiting the type IV genotype FimA (Mendez et al., 2019). The divergence between both lineages may be associated with differences in the virulence of the *P. gingivalis* strains, as a signal of a more specific adaptation to human host as a disrupting virulence factor. On the other hand, hemaglutinins (HagA or HagB) are well conserved in the trifecta, as a virulence factor important for heme acquisition, periodontal inflammation, and tissue breakdown. The evolutionary pattern of those hemaglutinins is almost completely species-specific, suggesting a strong vertical transference pattern, and potentially, another host-specific adaptation.

5. Conclusion

The following study, aimed to understand the evolution of *P. gingivalis* in the context of the *Porphyromonas* genus, could confirm the high relatedness with *P. gulae* but also found an unexpected relatedness with the marsupial-associated *P. loveana*; in coincidence with this relationship, several marker genes are more conserved in this group in comparison with the rest of Porphyromonads, suggesting that this group of three organisms was involved in the origin of the phenotype causing periodontal disease. Since those organisms have a characteristic presence in mammals, periodontal disease could be as old as the origin of the mammals.

The phylogenomic approach also found an interesting set of described and undescribed species, offering new opportunities for posterior studies. The phylogenetic analysis of the markers also opens the opportunity to study the properties of those virulence factors in the non-periodontal *Porphyromonas* species.

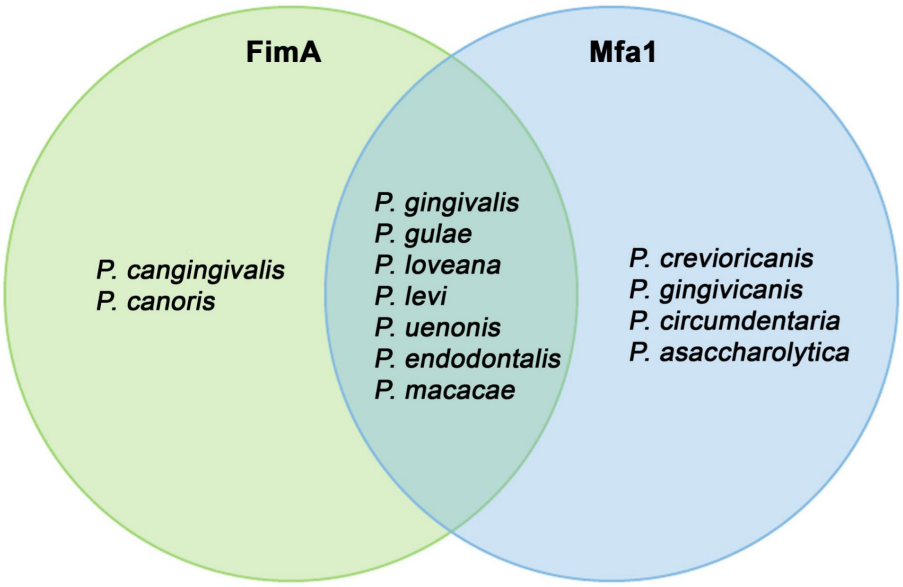


FIGURE 10
Organisms with FimA and MfaA observed in the *Porphyromonas* genome dataset. This Venn diagram showed which organisms contained encoded both phylogenetically-verified FimA and MfaA, and which organism contains Mfa1 or FimA only.

Data availability statement

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

Author contributions

JC and CC conceived the study. MM-O, JN-B, DG, and JC analyzed the data. MM-O, JN-B, DG, EV, JR-P, CC, and JC collaboratively elaborated, edited, and corrected the text and figures in the manuscript. All authors contributed to the article and approved the submitted version.

Funding

JC is supported by ANID Fondecyt project 11200209, CC is supported by ANID Fondecyt project 11190073, and JR-P is supported by ANID Fondecyt project 1221064. JN-B is supported by the Integrative Genomics Ph.D. program (Universidad Mayor, Chile).

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.2023.1226166/full#supplementary-material

References

Aleksijević, L. H., Aleksijević, M., Škrlec, I., Šram, M., Šram, M., and Talapko, J. (2022). *Porphyromonas gingivalis* virulence factors and clinical significance in periodontal disease and coronary artery diseases. *Pathogens* 11:1173. doi: 10.3390/pathogens11101173

Arjunan, P., El-Awady, A., Dannebaum, R. O., Kunde-Ramamoorthy, G., and Cutler, C. W. (2016). High-throughput sequencing reveals key genes and immune homeostatic pathways activated in myeloid dendritic cells by *Porphyromonas gingivalis* 381 and its fimbrial mutants. *Mol Oral Microbiol* 31, 78–93. doi: 10.1111/omi.12131

Bignell, C., and Thomas, C. M. (2001). The bacterial ParA-ParB partitioning proteins. *J. Biotechnol.* 91, 1–34. doi: 10.1016/s0168-1656(01)00293-0

Bird, P. S., Trott, D. J., Mikkelsen, D., Milinovich, G. J., Hillman, K. M., Burrell, P. C., et al. (2016). *Porphyromonas loveana* sp. nov., isolated from the oral cavity of Australian marsupials. *Int. J. Syst. Evol. Microbiol.* 66, 3771–3778. doi: 10.1099/ijsem.0.000898

Bosi, E., Fani, R., and Fondi, M. (2015). Defining orthologs and pangenome size metrics. *Methods Mol. Biol.* 1231, 191–202. doi: 10.1007/978-1-4939-1720-4_13

Bostanci, N., and Belibasakis, G. N. (2012). *Porphyromonas gingivalis*: an invasive and evasive opportunistic oral pathogen. *FEMS Microbiol. Lett.* 333, 1–9. doi: 10.1111/j.1574-6968.2012.02579.x

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome

(MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi: 10.1038/nbt.3893

Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A., and MacLean, C. (2019). The ecology and evolution of pangenomes. *Curr. Biol.* 29, R1094–R1103. doi: 10.1016/j.cub.2019.08.012

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

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., et al. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15, 1553–1565. doi: 10.1101/gr.4326505

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

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

Chen, W. A., Dou, Y., Fletcher, H. M., and Boskovic, D. S. (2023). Local and systemic effects of *Porphyromonas gingivalis* infection. *Microorganisms* 11, 470. doi: 10.3390/microorganisms11020470

Chen, T., Hosogi, Y., Nishikawa, K., Abbey, K., Fleischmann, R. D., Walling, J., et al. (2004). Comparative whole-genome analysis of virulent and avirulent strains of *Porphyromonas gingivalis*. *J. Bacteriol.* 186, 5473–5479. doi: 10.1128/JB.186.16.5473-5479.2004

Chen, T., Siddiqui, H., and Olsen, I. (2017). In silico comparison of 19 *Porphyromonas gingivalis* strains in genomics, phylogenetics, Phylogenomics and functional genomics. *Front. Cell. Infect. Microbiol.* 7:28. doi: 10.3389/fcimb.2017.00028

Chung, W. O., Demuth, D. R., and Lamont, R. J. (2000). Identification of a *Porphyromonas gingivalis* receptor for the *Streptococcus gordonii* SspB protein. *Infect. Immun.* 68, 6758–6762. doi: 10.1128/IAI.68.12.6758-6762.2000

Cifelli, R. L., and Davis, B. M. (2003). Paleontology. Marsupial origins. *Science* 302, 1899–1900. doi: 10.1126/science.1092272

Collins, M. D., Love, D. N., Karjalainen, J., Kanervo, A., Forsblom, B., Willems, A., et al. (1994). Phylogenetic analysis of members of the genus Porphyromonas and description of *Porphyromonas cangingivalis* sp. nov. and *Porphyromonas cansulci* sp. nov. *Int. J. Syst. Bacteriol.* 44, 674–679. doi: 10.1099/00207713-44-4-674

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912. doi: 10.1093/bioinformatics/btq315

Dashper, S. G., Mitchell, H. L., Seers, C. A., Gladman, S. L., Seemann, T., Bulach, D. M., et al. (2017). *Porphyromonas gingivalis* uses specific domain rearrangements and allelic exchange to generate diversity in surface virulence factors. *Front. Microbiol.* 8:48. doi: 10.3389/fmicb.2017.00048

Diard, M., and Hardt, W.-D. (2017). Evolution of bacterial virulence. *FEMS Microbiol. Rev.* 41, 679–697. doi: 10.1093/femsre/fux023

Eaton, D. A. R. (2020). Toytree: a minimalist tree visualization and manipulation library for python. *Methods Ecol. Evol.* 11, 187–191. doi: 10.1111/2041-210X.13313

El-Awady, A. R., Miles, B., Scisci, E., Kurago, Z. B., Palani, C. D., Arce, R. M., et al. (2015). *Porphyromonas gingivalis* evasion of autophagy and intracellular killing by human myeloid dendritic cells involves DC-SIGN-TLR2 crosstalk. *PLoS Pathog.* 10:e1004647. doi: 10.1371/journal.ppat.1004647

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

Endo, A., Watanabe, T., Ogata, N., Nozawa, T., Aikawa, C., Arakawa, S., et al. (2015). Comparative genome analysis and identification of competitive and cooperative interactions in a polymicrobial disease. *ISME J.* 9, 629–642. doi: 10.1038/ismej.2014.155

Eskan, M. A., Hajishengallis, G., and Kinane, D. F. (2007). Differential activation of human gingival epithelial cells and monocytes by *Porphyromonas gingivalis* fimbriae. *Infect. Immun.* 75, 892–898. doi: 10.1128/IAI.01604-06

Finegold, S. M., Vaisanen, M.-L., Rautio, M., Eerola, E., Summanen, P., Molitoris, D., et al. (2004). *Porphyromonas uenonis* sp. nov., a pathogen for humans distinct from *P. asaccharolytica* and *P. endodontalis*. *J. Clin. Microbiol.* 42, 5298–5301. doi: 10.1128/JCM.42.11.5298-5301.2004

Fournier, D., Mouton, C., Lapierre, P., Kato, T., Okuda, K., and Ménard, C. (2001). *Porphyromonas gulae* sp. nov., an anaerobic, gram-negative coccobacillus from the gingival sulcus of various animal hosts. *Int. J. Syst. Evol. Microbiol.* 51, 1179–1189. doi: 10.1099/00207713-51-3-1179

Fujiwara-Takahashi, K., Watanabe, T., Shimogishi, M., Shibasaki, M., Umeda, M., Izumi, Y., et al. (2020). Phylogenetic diversity in FIM and MFA gene clusters between *Porphyromonas gingivalis* and *Porphyromonas gulae*, as a potential cause of host specificity. *J. Oral Microbiol.* 12:1775333. doi: 10.1080/20002297.2020.1775333

Gaba, S., Kumari, A., Medema, M., and Kaushik, R. (2020). Pan-genome analysis and ancestral state reconstruction of class halobacteria: probability of a new super-order. *Sci. Rep.* 10:21205. doi: 10.1038/s41598-020-77723-6

Gabarrini, G., Chlebowicz, M. A., Vega Quiroz, M. E., Veloo, A. C. M., Rossen, J. W. A., Harmsen, H. J. M., et al. (2018). Conserved citrullinating exoenzymes in *Porphyromonas* species. *J. Dent. Res.* 97, 556–562. doi: 10.1177/0022034517747575

Gabarrini, G., de Smit, M., Westra, J., Brouwer, E., Vissink, A., Zhou, K., et al. (2015). The peptidylarginine deiminase gene is a conserved feature of *Porphyromonas gingivalis*. *Sci. Rep.* 5:13936. doi: 10.1038/srep13936

Gerhardt, A., Cinkaya, I., Linder, D., Huisman, G., and Buckel, W. (2000). Fermentation of 4-aminobutyrate by clostridium aminobutyricum: cloning of two genes involved in the formation and dehydration of 4-hydroxybutyryl-CoA. *Arch. Microbiol.* 174, 189–199. doi: 10.1007/s002030000195

Gibson, F. C., and Genco, C. A. (2006). "The Genus Porphyromonas," In *The Prokaryotes*, (Eds.) M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer and E. Stackebrandt (New York, NY: Springer New York), 428–454.

Guilloux, C.-A., Lamoureux, C., Beauruelle, C., and Héry-Arnaud, G. (2021). Porphyromonas: a neglected potential key genus in human microbiomes. *Anaerobe* 68:102230. doi: 10.1016/j.anaerobe.2020.102230

Guo, Y., Nguyen, K.-A., and Potempa, J. (2010). Dichotomy of gingipains action as virulence factors: from cleaving substrates with the precision of a surgeon's knife to a meat chopper-like brutal degradation of proteins. *Periodontol.* 54, 15–44. doi: 10.1111/j.1600-0757.2010.00377.x

Gupta, R. S., Patel, S., Saini, N., and Chen, S. (2020). Robust demarcation of 17 distinct bacillus species clades, proposed as novel *Bacillaceae genera*, by phylogenomics and comparative genomic analyses: description of *Robertmurraya kyonggiensis* sp. nov. and proposal for an emended genus bacillus limiting it only to the members of the subtilis and cereus clades of species. *Int. J. Syst. Evol. Microbiol.* 70, 5753–5798. doi: 10.1099/ijsem.0.004475

Hajishengallis, G., and Lamont, R. J. (2014). Breaking bad: manipulation of the host response by *Porphyromonas gingivalis*. *Eur. J. Immunol.* 44, 328–338. doi: 10.1002/eji.201344202

Hajishengallis, G., and Lamont, R. J. (2016). Dancing with the stars: how choreographed bacterial interactions dictate nososymbiocity and give rise to keystone pathogens, accessory pathogens, and pathobionts. *Trends Microbiol.* 24, 477–489. doi: 10.1016/j.tim.2016.02.010

Hajishengallis, G., Tapping, R. I., Harokopakis, E., Nishiyama, S., Ratti, P., Schifferle, R. E., et al. (2006). Differential interactions of fimbriae and lipopolysaccharide from *Porphyromonas gingivalis* with the toll-like receptor 2-centred pattern recognition apparatus. *Cell. Microbiol.* 8, 1557–1570. doi: 10.1111/j.1462-5822.2006.00730.x

Hamada, N., Takahashi, Y., Watanabe, K., Kumada, H., Oishi, Y., and Umemoto, T. (2008). Molecular and antigenic similarities of the fimbrial major components between *Porphyromonas gulae* and *P. gingivalis*. *Vet. Microbiol.* 128, 108–117. doi: 10.1016/j.vetmic.2007.09.014

Harokopakis, E., Albzreh, M. H., Martin, M. H., and Hajishengallis, G. (2006). TLR2 transmodulates monocyte adhesion and transmigration via Rac1- and PI3K-mediated inside-out signaling in response to *Porphyromonas gingivalis* fimbriae. *J. Immunol.* 176, 7645–7656. doi: 10.4049/jimmunol.176.12.7645

Hasegawa, Y., and Nagano, K. (2021). *Porphyromonas gingivalis* FimA and Mfa1 fimbriae: current insights on localization, function, biogenesis, and genotype. *Jpn. Dent Sci Rev* 57, 190–200. doi: 10.1016/j.jdsr.2021.09.003

Henry, L. G., McKenzie, R. M. E., Robles, A., and Fletcher, H. M. (2012). Oxidative stress resistance in *Porphyromonas gingivalis*. *Future Microbiol.* 7, 497–512. doi: 10.2217/fmb.12.17

Hirasawa, M., and Takada, K. (1994). *Porphyromonas gingivicanis* sp. nov. and *Porphyromonas crevioricanis* sp. nov., isolated from beagles. *Int. J. Syst. Bacteriol.* 44, 637–640. doi: 10.1099/00207713-44-4-637

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Hyun, J. C., Monk, J. M., and Palsson, B. O. (2022). Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* 23:7. doi: 10.1186/s12864-021-08223-8

Imirzalioglu, C., Sethi, S., Schneider, C., Hain, T., Chakraborty, T., Mayser, P., et al. (2014). Distinct polymicrobial populations in a chronic foot ulcer with implications for diagnostics and anti-infective therapy. *BMC. Res. Notes* 7:196. doi: 10.1186/1756-0500-7-196

Inaba, H., Nomura, R., Kato, Y., Takeuchi, H., Amano, A., Asai, F., et al. (2019). Adhesion and invasion of gingival epithelial cells by *Porphyromonas gulae*. *PLoS One* 14:e0213309. doi: 10.1371/journal.pone.0213309

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

Iwashita, N., Nomura, R., Shirai, M., Kato, Y., Murakami, M., Matayoshi, S., et al. (2019). Identification and molecular characterization of *Porphyromonas gulae* fimA types among cat isolates. *Vet. Microbiol.* 229, 100–109. doi: 10.1016/j.vetmic.2018.12.018

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, L., Han, N., Du, J., Guo, L., Luo, Z., and Liu, Y. (2019). Pathogenesis of important virulence factors of *Porphyromonas gingivalis* via toll-like receptors. *Front. Cell. Infect. Microbiol.* 9:262. doi: 10.3389/fcimb.2019.00262

Kamp, H. D., Patimalla-Dipali, B., Lazinski, D. W., Wallace-Gadsden, F., and Camilli, A. (2013). Gene fitness landscapes of *Vibrio cholerae* at important stages of its life cycle. *PLoS Pathog.* 9:e1003800. doi: 10.1371/journal.ppat.1003800

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformatics* 20, 1160–1166. doi: 10.1093/bib/bbx108

Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* 14:289. doi: 10.1186/1471-2105-14-289

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLoS Genet.* 8:e1002787. doi: 10.1371/journal.pgen.1002787

Kuboniwa, M., Amano, A., Hashino, E., Yamamoto, Y., Inaba, H., Hamada, N., et al. (2009). Distinct roles of long/short fimbriae and gingipains in homotypic biofilm development by *Porphyromonas gingivalis*. *BMC Microbiol.* 9:105. doi: 10.1186/1471-2180-9-105

Laland, K. N., Odling-Smee, F. J., and Feldman, M. W. (1999). Evolutionary consequences of niche construction and their implications for ecology. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10242–10247. doi: 10.1073/pnas.96.18.10242

Lasica, A. M., Ksiazek, M., Madej, M., and Potempa, J. (2017). The type IX secretion system (T9SS): highlights and recent insights into its structure and function. *Front. Cell. Infect. Microbiol.* 7:215. doi: 10.3389/fcimb.2017.00215

Lenzo, J. C., O'Brien-Simpson, N. M., Orth, R. K., Mitchell, H. L., Dashper, S. G., and Reynolds, E. C. (2016). *Porphyromonas gulae* has virulence and immunological characteristics similar to those of the human periodontal pathogen *Porphyromonas gingivalis*. *Infect. Immun.* 84, 2575–2585. doi: 10.1128/IAI.01500-15

Lewis, J. P., Yanamandra, S. S., and Anaya-Bergman, C. (2012). HcpR of *Porphyromonas gingivalis* is required for growth under nitrosative stress and survival within host cells. *Infect. Immun.* 80, 3319–3331. doi: 10.1128/IAI.00561-12

Lithgow, K. V., Buchholz, V. C. H., Ku, E., Konschuh, S., D'Aubeterre, A., and Sycuro, L. K. (2022). Protease activities of vaginal *Porphyromonas* species disrupt coagulation and extracellular matrix in the cervicovaginal niche. *NPJ Biofilms Microbiomes* 8:8. doi: 10.1038/s41522-022-00270-7

Liu, J., Zhang, Y., Lei, X., and Zhang, Z. (2008). Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol.* 9:R69. doi: 10.1186/gb-2008-9-4-r69

Lombardo Bedran, T. B., Marcantonio, R. A. C., Spin Neto, R., Alves Mayer, M. P., Grenier, D., Spolidorio, L. C., et al. (2012). *Porphyromonas endodontalis* in chronic periodontitis: a clinical and microbiological cross-sectional study. *J. Oral Microbiol.* 4:10123. doi: 10.3402/jom.v4i0.10123

Lourbakos, A., Yuan, Y. P., Jenkins, A. L., Travis, J., Andrade-Gordon, P., Santulli, R., et al. (2001). Activation of protease-activated receptors by gingipains from *Porphyromonas gingivalis* leads to platelet aggregation: a new trait in microbial pathogenicity. *Blood* 97, 3790–3797. doi: 10.1182/blood.v97.12.3790

Love, D. N., Bailey, G. D., Collings, S., and Briscoe, D. A. (1992). Description of *Porphyromonas circumdentaria* sp. nov. and reassignment of *Bacteroides salivosus* (Love, Johnson, Jones, and Calverley 1987) as Porphyromonas (Shah and Collins 1988) salivosa comb. nov. *Int. J. Syst. Bacteriol.* 42, 434–438. doi: 10.1099/00207713-42-3-434

Love, D. N., Karjalainen, J., Kanervo, A., Forsblom, B., Sarkiala, E., Bailey, G. D., et al. (1994). *Porphyromonas canoris* sp. nov., an asaccharolytic, black-pigmented species from the gingival sulcus of dogs. *Int. J. Syst. Bacteriol.* 44, 204–208. doi: 10.1099/00207713-44-2-204

Lunar Silva, I., and Cascales, E. (2021). Molecular strategies underlying *Porphyromonas gingivalis* virulence. *J. Mol. Biol.* 433:166836. doi: 10.1016/j.jmb.2021.166836

Madej, M., White, J. B. R., Nowakowska, Z., Rawson, S., Scavenius, C., Enghild, J. J., et al. (2020). Structural and functional insights into oligopeptide acquisition by the RagAB transporter from *Porphyromonas gingivalis*. *Nat. Microbiol.* 5, 1016–1025. doi: 10.1038/s41564-020-0716-y

Maresz, K. J., Hellvard, A., Sroka, A., Adamowicz, K., Bielecka, E., Koziel, J., et al. (2013). *Porphyromonas gingivalis* facilitates the development and progression of destructive arthritis through its unique bacterial peptidylarginine deiminase (PAD). *PLoS Pathog.* 9:e1003627. doi: 10.1371/journal.ppat.1003627

Matrishin, C. B., Haase, E. M., Dewhirst, F. E., Mark Welch, J. L., Miranda-Sanchez, F., MacFarland, D. C., et al. (2022). Phages are important unrecognized players in the ecology of the oral pathogen *Porphyromonas gingivalis*. *BioRxiv*, 12:30.519816. doi: 10.1101/2022.12.30.519816

Maturana, J. L., and Cárdenas, J. P. (2021). Insights on the evolutionary genomics of the Blautia genus: potential new species and genetic content among lineages. *Front. Microbiol.* 12:660920. doi: 10.3389/fmicb.2021.660920

Mei, M. L., Yan, Z., Duangthip, D., Niu, J. Y., Yu, O. Y., You, M., et al. (2020). Effect of silver diamine fluoride on plaque microbiome in children. *J. Dent.* 102:103479. doi: 10.1016/j.jdent.2020.103479

Mendez, K. N., Hoare, A., Soto, C., Bugueño, I., Olivera, M., Meneses, C., et al. (2019). Variability in genomic and virulent properties of *Porphyromonas gingivalis* strains isolated from healthy and severe chronic periodontitis individuals. *Front. Cell. Infect. Microbiol.* 9:246. doi: 10.3389/fcimb.2019.00246

Miller, D. P., Hutcherson, J. A., Wang, Y., Nowakowska, Z. M., Potempa, J., Yoder-Himes, D. R., et al. (2017). Genes contributing to *Porphyromonas gingivalis* fitness in abscess and epithelial cell colonization environments. *Front. Cell. Infect. Microbiol.* 7:378. doi: 10.3389/fcimb.2017.00378

Mishra, D., and Srinivasan, R. (2022). Catching a walker in the act—DNA partitioning by ParA family of proteins. *Front. Microbiol.* 13:856547. doi: 10.3389/fmicb.2022.856547

Mohanty, R., Asopa, S. J., Joseph, M. D., Singh, B., Rajguru, J. P., Saidath, K., et al. (2019). Red complex: polymicrobial conglomerate in oral flora: a review. *J. Family Med. Prim. Care* 8, 3480–3486. doi: 10.4103/jfmpc.jfmpc_759_19

Moldovan, M. A., and Gelfand, M. S. (2018). Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. *Front. Microbiol.* 9:428. doi: 10.3389/fmicb.2018.00428

Mysak, J., Podzimek, S., Sommerova, P., Lyuya-Mi, Y., Bartova, J., Janatova, T., et al. (2014). *Porphyromonas gingivalis*: major periodontopathic pathogen overview. *J Immunol Res* 2014:476068. doi: 10.1155/2014/476068

Nagano, K., Hasegawa, Y., Iijima, Y., Kikuchi, T., and Mitani, A. (2018). Distribution of *Porphyromonas gingivalis* fimA and mfa1 fimbrial genotypes in subgingival plaques. *PeerJ* 6:e5581. doi: 10.7717/peerj.5581

Naito, M., Hirakawa, H., Yamashita, A., Ohara, N., Shoji, M., Yukitake, H., et al. (2008). Determination of the genome sequence of *Porphyromonas gingivalis* strain ATCC 33277 and genomic comparison with strain W83 revealed extensive genome rearrangements in *P. gingivalis*. *DNA Res.* 15, 215–225. doi: 10.1093/dnares/dsn013

Nakayama, M., and Ohara, N. (2017). Molecular mechanisms of *Porphyromonas gingivalis* -host cell interaction on periodontal diseases. *Jpn. Dent. Sci. Rev.* 53, 134–140. doi: 10.1016/j.jdsr.2017.06.001

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

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* 86, 641–647. doi: 10.1046/j.1365-2540.2001.00895.x

Nuñez-Belmar, J., Morales-Olavarria, M., Vicencio, E., Vernal, R., Cárdenas, J. P., and Cortez, C. (2022). Contribution of -omics technologies in the study of *Porphyromonas gingivalis* during periodontitis pathogenesis: a minireview. *Int. J. Mol. Sci.* 24:620. doi: 10.3390/ijms24010620

O'Flynn, C., Deusch, O., Darling, A. E., Eisen, J. A., Wallis, C., Davis, I. J., et al. (2015). Comparative genomics of the genus Porphyromonas identifies adaptations for Heme synthesis within the prevalent canine Oral species *Porphyromonas cangingivalis*. *Genome Biol. Evol.* 7, 3397–3413. doi: 10.1093/gbe/evv220

Olsen, I., Chen, T., and Tribble, G. D. (2018). Genetic exchange and reassignment in *Porphyromonas gingivalis*. *J. Oral Microbiol.* 10:1457373. doi: 10.1080/20002297.2018.1457373

Olsen, I., Lambris, J. D., and Hajishengallis, G. (2017). *Porphyromonas gingivalis* disturbs host-commensal homeostasis by changing complement function. *J. Oral Microbiol.* 9:1340085. doi: 10.1080/20002297.2017.1340085

Park, K.-H. (2015). Roles of enzymes in glycogen metabolism and degradation in *Escherichia coli*. *J. Appl. Glycosci.* 62, 37–45. doi: 10.5458/jag.jag.JAG-2015_005

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

Paster, B. J., Dewhirst, F. E., Olsen, I., and Fraser, G. J. (1994). Phylogeny of *Bacteroides*, *Prevotella*, and *Porphyromonas* spp. and related bacteria. *J. Bacteriol.* 176, 725–732. doi: 10.1128/jb.176.3.725-732.1994

Peery, M. Z., Kirby, R., Reid, B. N., Stoelting, R., Doucet-Bëer, E., Robinson, S., et al. (2012). Reliability of genetic bottleneck tests for detecting recent population declines. *Mol. Ecol.* 21, 3403–3418. doi: 10.1111/j.1365-294X.2012.05635.x

Rajilić-Stojanović, M., and de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* 38, 996–1047. doi: 10.1111/1574-6976.12075

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

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226–235. doi: 10.1016/j.jtbi.2005.08.037

Roky, M., Trent, J. O., and Demuth, D. R. (2020). Identification of functional domains of the minor fimbrial antigen involved in the interaction of *Porphyromonas gingivalis* with oral streptococci. *Mol Oral Microbiol* 35, 66–77. doi: 10.1111/omi.12280

Romero-Lastra, P., Sánchez, M. C., Ribeiro-Vidal, H., Llama-Palacios, A., Figuero, E., Herrera, D., et al. (2017). Comparative gene expression analysis of *Porphyromonas gingivalis* ATCC 33277 in planktonic and biofilms states. *PLoS One* 12:e0174669. doi: 10.1371/journal.pone.0174669

Rouli, L., Merhej, V., Fournier, P. E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* 7, 72–85. doi: 10.1016/j.nmni.2015.06.005

Sakamoto, M., Li, D., Shibata, Y., Takeshita, T., Yamashita, Y., and Ohkuma, M. (2015). *Porphyromonas pasteri* sp. nov., isolated from human saliva. *Int. J. Syst. Evol. Microbiol.* 65, 2511–2515. doi: 10.1099/ijs.0.000294

Sakamoto, M., and Ohkuma, M. (2013). *Porphyromonas crevioricanis* is an earlier heterotypic synonym of *Porphyromonas cansulci* and has priority. *Int. J. Syst. Evol. Microbiol.* 63, 454–457. doi: 10.1099/ijs.0.042531-0

Sato, T., Tomida, J., Naka, T., Fujiwara, N., Hasegawa, A., Hoshikawa, Y., et al. (2015). *Porphyromonas bronchialis* sp. nov. isolated from intraoperative bronchial fluids of a patient with non-small cell lung cancer. *Tohoku J. Exp. Med.* 237, 31–37. doi: 10.1620/tjem.237.31

Sawana, A., Adeolu, M., and Gupta, R. S. (2014). Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus *Paraburkholderia* gen. Nov. harboring environmental species. *Front. Genet.* 5:429. doi: 10.3389/fgene.2014.00429

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

Shah, H. N., and Collins, M. D. (1988). Proposal for reclassification of *Bacteroides asaccharolyticus*, Bacteroides gingivalis, and *Bacteroides endodontalis* in a new genus. *Porphyromonas. Int. J. Syst. Bacteriol.* 38, 128–131. doi: 10.1099/00207713-38-1-128

Summanen, P., and Finegold, S. M. (2015). "*Porphyromonas*" in *Bergey's Manual of Systematics of Archaea and Bacteria*. eds. W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun and P. DeVoset al. (Wiley Online Library), 1–14.

Summanen, P. H., Lawson, P. A., and Finegold, S. M. (2009). *Porphyromonas bennonis* sp. nov., isolated from human clinical specimens. *Int. J. Syst. Evol. Microbiol.* 59, 1727–1732. doi: 10.1099/ijs.0.001909-0

Suzuki, Y., Kikuchi, T., Goto, H., Takayanagi, Y., Kawamura, S., Sawada, N., et al. (2022). *Porphyromonas gingivalis* fimbriae induce Osteoclastogenesis via toll-like receptors in RAW264 cells. *Int. J. Mol. Sci.* 23:15293. doi: 10.3390/ijms232315293

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.* 804, 281–295. doi: 10.1007/978-1-61779-361-5_15

Veronika, M., Welsch, R., Ng, A., Matsudaira, P., and Rajapakse, J. C. (2011). Correlation of cell membrane dynamics and cell motility. *BMC Bioinformatics* 12:S19. doi: 10.1186/1471-2105-12-S13-S19

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi: 10.1146/annurev-genet-111212-133526

Waller, T., Kesper, L., Hirschfeld, J., Dommisch, H., Kölpin, J., Oldenburg, J., et al. (2016). *Porphyromonas gingivalis* outer membrane vesicles induce selective tumor necrosis factor tolerance in a toll-like receptor 4- and mTOR-dependent manner. *Infect. Immun.* 84, 1194–1204. doi: 10.1128/IAI.01390-15

Wang, H., Zhang, W., Wang, W., and Zhang, L. (2020). The prevalence of fimA genotypes of *Porphyromonas gingivalis* in patients with chronic periodontitis: a meta-analysis. *PLoS One* 15:e0240251. doi: 10.1371/journal.pone.0240251

Watanabe, T., Shibasaki, M., Maruyama, F., Sekizaki, T., and Nakagawa, I. (2017). Investigation of potential targets of *Porphyromonas* CRISPRs among the genomes of *Porphyromonas* species. *PLoS One* 12:e0183752. doi: 10.1371/journal.pone.0183752

Yamada, M., Ikegami, A., and Kuramitsu, H. K. (2005). Synergistic biofilm formation by treponema denticola and *Porphyromonas gingivalis*. *FEMS Microbiol. Lett.* 250, 271–277. doi: 10.1016/j.femsle.2005.07.019

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

Zamora-Cintas, M., Marín, M., Quiroga, L., Martínez, A., Fernández-Chico, M. A., Bouza, E., et al. (2018). Identification of *Porphyromonas* isolates from clinical origin using MALDI-TOF mass spectrometry. *Anaerobe* 54, 197–200. doi: 10.1016/j.anaerobe.2018.06.017

Zenobia, C., and Hajishengallis, G. (2015). *Porphyromonas gingivalis* virulence factors involved in subversion of leukocytes and microbial dysbiosis. *Virulence* 6, 236–243. doi: 10.1080/21505594.2014.999567

Zhang, D., Li, S., Hu, L., Sheng, L., and Chen, L. (2015). Modulation of protease-activated receptor expression by *Porphyromonas gingivalis* in human gingival epithelial cells. *BMC Oral Health* 15:128. doi: 10.1186/s12903-015-0105-8

Zhang, Q., Tyler-Smith, C., and Long, Q. (2015). An extended Tajima's D neutrality test incorporating SNP calling and imputation uncertainties. *Stat. Interface* 8, 447–456. doi: 10.4310/SII.2015.v8.n4.a4

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 August 2023
DOI 10.3389/fmicb.2023.1207878

Check for updates

OPEN ACCESS

EDITED BY
Paulo Jorge Dias,
University of Lisbon, Portugal

REVIEWED BY
Abbas Yadegar,
Shahid Beheshti University of Medical Sciences, Iran
Amin Talebi Bezmin Abadi,
Tarbiat Modares University, Iran
Sílvia A. Sousa,
Institute for Bioengineering and Biosciences, Portugal

*CORRESPONDENCE
Haiying Gu
✉ guhaiying@nbu.edu.cn

†These authors have contributed equally to this work

RECEIVED 18 April 2023
ACCEPTED 25 July 2023
PUBLISHED 08 August 2023

CITATION
Zhang Y, Gu H, Shi Z, Chen W, Li A, Ye W, Zhang C, Yuan H and Zhao M (2023) High prevalence of *Helicobacter pylori* mixed infections identified by multilocus sequence typing in Ningbo, China.
Front. Microbiol. 14:1207878.
doi: 10.3389/fmicb.2023.1207878

COPYRIGHT
© 2023 Zhang, Gu, Shi, Chen, Li, Ye, Zhang, Yuan and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

High prevalence of *Helicobacter pylori* mixed infections identified by multilocus sequence typing in Ningbo, China

Yongxiong Zhang[1†], Haiying Gu[1,2*†], Zhouhong Shi[3], Weiqin Chen[3], Airu Li[2], Weiwei Ye[3], Cheng Zhang[3], Huikun Yuan[2] and Mingming Zhao[1]

[1]Health Science Center, Ningbo University, Ningbo, Zhejiang, China, [2]Laboratory of Gastroenterology, The Affiliated Hospital of Medical School, Ningbo University, Ningbo, Zhejiang, China, [3]Department of Gastrology, Ninghai First Hospital, Ningbo, Zhejiang, China

This study used multilocus sequence typing (MLST) to investigate the prevalence of *Helicobacter pylori* (*H. pylori*) mixed infections and *H. pylori* mixed infections involving unrelated strains; and determined the phylogeographic groups of *H. pylori* recovered from patients in Ningbo, China. A total of 156 *H. pylori* isolates were obtained from a convenience sample of 33 patients with culture-positive *H. pylori* infection. MLST was used to classify 150 *H. pylori* clinical isolates and 12 methodological control strains (6 clinical isolates and 6 strains of American Type Culture Collection *H. pylori*) into 43 and 12 sequence types (STs), respectively. In this study, 246 new alleles and 53 new STs were identified by MLST. The prevalence of mixed infections was 41% (11/27). The prevalence of *H. pylori* mixed infections involving unrelated strains was 46% (5/11) and the prevalence of *H. pylori* mixed infections involving completely unrelated strains (strains with all 7 housekeeping genes different) was 36% (4/11). A phylogenetic tree was created to determine the evolutionary relationships between different strains. The STs in this study were clustered within the hspEAsia subgroup (98%) and hpEurope group (2%). *H. pylori* mixed infections were common in Ningbo, China. The *H. pylori* isolates belonging to the hpEurope group were recovered from three different biopsy samples in a native Chinese patient. Most of *H. pylori* strains colonizing the antrum, corpus, and duodenum bulb were homologous.

KEYWORDS

***Helicobacter pylori*, mixed infections, unrelated strains, evolutionary relationships, homologous, multilocus sequence typing, phylogeographic groups**

Abbreviations: ATCC, American Type Culture Collection; *atpA*, gene encoding an ATP synthase alpha chain; *efp*, gene encoding an elongation factor P; h, number of haplotypes; Hd, haplotype diversity; MLST, multilocus sequence typing; *mutY*, gene encoding a DNA glycosylase; PCR, polymerase chain reaction; Pi, nucleotide diversity; *ppa*, gene encoding an inorganic pyrophosphatase; RAPD, random amplified polymorphism DNA; S, number of polymorphic sites; ST, sequence type; TLV, triple locus variants; *trpC*, gene encoding an anthranilate isomerase; *urel*, gene encoding a urease subunit I; *yphC*, gene encoding a GTPase.

Introduction

Helicobacter pylori (*H. pylori*) is a human pathogen that colonizes the gastrointestinal mucosa and causes chronic active gastritis, peptic ulcers, lymphoid tissue lymphoma, and gastric adenocarcinoma (Blaser and Atherton, 2004; Burkitt et al., 2017; Camilo et al., 2017). It is a gram-negative microaerophilic bacterium and has a high prevalence worldwide.

The gastrointestinal mucosa can be colonized by single or multiple *H. pylori* strains (Blaser and Berg, 2001; Kusters et al., 2006; Talebi Bezmin Abadi and Perez-Perez, 2016). *H. pylori* mixed infections and heteroresistance may lead to the failure of eradication treatment, based on different resistance of isolates from the antrum and corpus to antimicrobial agents (Farzi et al., 2019). *H. pylori* mixed infections are a cause for concern. The prevalence of *H. pylori* mixed infections varies from 0 to 100% in different geographic regions (Ben Mansour et al., 2016). The evolutionary relationships between different *H. pylori* isolates recovered from individual hosts are unclear. Previous studies have described four patterns of mixed infection: (i) most *H. pylori* isolates are related and exhibit slightly different patterns, and only few isolates are unrelated (Kim et al., 2003; Carroll et al., 2004; Kao et al., 2014; Farzi et al., 2015; Mendoza-Elizalde et al., 2019); (ii) all isolates are related (Patra et al., 2012; Ren et al., 2012; Kibria et al., 2015); (iii) isolates are predominantly unrelated (70%), with a few (30%) related isolates exhibiting slightly different patterns (Seo et al., 2019); (iv) all isolates are unrelated or show two independent populations of *H. pylori* (Carroll et al., 2004; Matteo et al., 2007; Salama et al., 2007). The prevalence of mixed infections varies according to the geographic region and the participant selection criteria (Kersulyte et al., 2000; Wong et al., 2001; Ben Mansour et al., 2016). *H. pylori* isolates with slightly different DNA fingerprinting patterns from a single host have revealed the mechanisms of microevolution (Marshall et al., 1998; Kim et al., 2003; Carroll et al., 2004; Mendoza-Elizalde et al., 2019). Due to its limitations, the DNA fingerprinting method cannot determine the evolutionary relationships of isolates from a single host (Carroll et al., 2004; Matteo et al., 2007).

Multilocus sequence typing (MLST) is based on comparison of the sequences of seven housekeeping genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, and *yphC*) (Osaki et al., 2015; Raaf et al., 2017; Mendoza-Elizalde et al., 2019) and has the advantages of high repeatability, high resolution, and normalization. Moreover, this method can be used to determine the evolutionary relationships of isolates recovered from a single patient with *H. pylori* mixed infections (Mendoza-Elizalde et al., 2019). The prevalence of *H. pylori* mixed infections involving strains that are unrelated according to MLST is unclear. Importantly, the geographical location of the human host can be inferred from the MLST results (Mégraud et al., 2016; Gutiérrez-Escobar et al., 2017; Vazirzadeh et al., 2022). However, there is a lack of data on the geographic grouping of *H. pylori* isolates from patients in southern urban China.

Therefore, the purpose of this study was to characterize *H. pylori* mixed infections by MLST. More specifically, we investigated the prevalence of *H. pylori* mixed infections involving unrelated strains and phylogeographic groups of *H. pylori* isolates recovered from single patients in Ningbo, China.

Materials and methods

Ethics statement

Ethical approval for this study was obtained from the Medical Research Committee of the Ningbo University School of Medicine (approval number: NBU-2021-129).

Patients and specimens

A convenience sample of 33 patients from the Ninghai First Hospital and the Affiliated Hospital of Ningbo University School of Medicine who were culture positive for *H. pylori* participated in the study. Patients were excluded if they had taken antibiotics, proton pump inhibitors, H2-receptor antagonists, non-steroidal anti-inflammatory drugs, or bismuth-containing compounds within the 15 days prior to endoscopy. All patients provided written informed consent to participate in the study. Six patients were assigned to the control group, in which the prevalence of *H. pylori* mixed infections was not compared. Two or three mucosal biopsy specimens were collected from the antrum, corpus, and/or duodenal bulb of each patient for bacterial culture. The specimens were immediately placed in a transport medium (Gu's Kit for Preservation of *Helicobacter pylori*, TianKuo, Ningbo Xunjian Biotechnology Co., Ltd., Ningbo, China) and transported to the laboratory at 2–8°C. Two other biopsy specimens were collected from each patient for histological examination.

Isolation and identification of *H. pylori*

The biopsy specimens were homogenized and inoculated on Gu's plates (Gu's Medium for Rapid Isolation of *Helicobacter pylori*, TianKuo, Ningbo Xunjian Biotechnology Co., Ltd., Ningbo, China; PCT WO2022178982A1). The plates were incubated at 37°C for 2–5 days in a microaerophilic environment (3–5% O_2, 5–10% CO_2, 5–10% H_2, 75–87% N_2) with 100% humidity (Anoxomat; MART Microbiology BV, Drachten, Netherlands). A pool of colonies was selected from the primary culture plates for subculturing. Single colonies were obtained from biopsy specimens of different sites of each patient using the colony suspension dilution method, and 1–10 single colonies were selected and passaged separately to obtain single colony isolates.

One of the three isolation schemes a, b, and c was selected to analyze *H. pylori* mixed infections in each patient, as follows: (a) One *H. pylori* colony was isolated from each of multiple biopsy specimens of a single patient; (b) Multiple (3–11) single colonies were obtained from one biopsy specimen of a single patient; (c) Multiple (2–10) single colonies were isolated from each of several biopsy specimens of a single patient.

Helicobacter pylori was identified based on colony morphology and Gram staining as a gullwing-shaped bacterium; positive reactions for catalase, oxidase, and urease (Gu's Kit for Rapid Identification of *Helicobacter pylori*, TianKuo, Ningbo Xunjian Biotechnology Co., Ltd., Ningbo, China); and *H. pylori* antigen testing (*H. pylori* Antigen Rapid Test, Abon Biopharm, Hangzhou, China). Six American Type Culture Collection (ATCC) *H. pylori* reference strains (ATCC 43629, ATCC 700392, ATCC 51932, ATCC 700824, ATCC 43579, and ATCC 49503) were used as controls.

DNA extraction

Genomic DNA was extracted using a HiPure Bacterial DNA Kit (Guangzhou Magen Biotechnology Co., Ltd., Guangzhou, China) according to the manufacturer's instructions. The extracted genomic DNA was stored at −20°C until being amplified using polymerase chain reaction (PCR).

Multilocus sequence typing

The extracted DNA was used as a template for PCR amplification of seven housekeeping genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, and *yphC*). The primers used for the MLST housekeeping genes are shown in **Table 1**. The PCR amplification conditions used were as described by Achtman et al. (1999). The PCR products were detected, purified, and sequenced using Sanger sequencing (QingKe Biotechnology Co., Ltd., Hangzhou, China). The sequence peaks were interpreted using Chromas software to remove miscellaneous peaks, and the resulting sequences were imported into DNA Star software and assembled using the SeqMan function. The trimmed sequences of the seven housekeeping genes were imported into the *H. pylori* PubMLST database[1] to identify allelic matches. The STs were identified based on the combination of MLST alleles. Nucleotide sequences that did not match existing PubMLST sequences were submitted to the database for new number assignment and sequence typing.

Phylogeny, genealogic analysis, and genetic diversity analysis

The genealogic relationships among *H. pylori* isolates from the patients with mixed infections were investigated using PHYLOViZ (Mendoza-Elizalde et al., 2019). The goeBURST algorithm was used to demonstrate the relationship between clonal complexes per patient at the triple locus variants (TLV) level. The neighbor-joining algorithm with the Hamming distance was utilized to calculate the phylogenetic distance among strains using the Saitou and Nei (1987) criterion.

A phylogenetic tree was constructed using the MEGA 11 software for evolutionary analysis (Tamura et al., 2021). The concatenated nucleotide sequences of seven housekeeping genes in the studied *H. pylori* clinical isolates and six reference strains (ATCC 43629, ATCC 700392, ATCC 51932, ATCC 700824, ATCC 43579, and ATCC 49503) were aligned in ClustalW. Phylogenetic distances were computed using the neighbor-joining method (Saitou and Nei, 1987) with the Kimura 2-parameter model of nucleotide substitution (Kimura, 1980). The reliability of clustering was assessed using a bootstrap test (1,000 bootstrap replications). Polymorphisms in the housekeeping genes of *H. pylori* isolates obtained from patients with mixed infections were analyzed using DnaSP v6, which included polymorphic sites (S), nucleotide diversity (Pi), number of haplotypes (h), and haplotype diversity (Hd).

In addition, we downloaded 284 concatenated nucleotide sequences of the seven housekeeping genes of *H. pylori* strains from the PubMLST database[2] as representatives of different geographical groups to determine the geographical type of *H. pylori* strains obtained in this study. After aligning the concatenated nucleotide sequences of the seven housekeeping genes in the studied strains (55 STs) and reference sequences (284 STs) by ClustalW, a phylogenetic tree was constructed in MEGA 11. The reference sequences for the geographical groups were as follows: hpEurope, 75 sequences; hpAsia2, 18 sequences; hspMaori, 35 sequences; hspEAsia, 50 sequences; hspAmerind, 9 sequences; hpAfrica2, 3 sequences; hspSAfrica, 25 sequences; hspWAfrica, 29 sequences; hpNEAfrica, 20 sequences; and hpSahul, 20 sequences.

1 http://pubmlst.org/helicobacter

2 https://pubmlst.org/organisms/helicobacter-pylori

TABLE 1 Primers for the multilocus sequence typing (MLST) housekeeping gene.

Gene	Primer sequence (5′→3′)	Product size (bp)	Annealing temperature
atpA	*atpA*-F: GGACTAGCGTTAAACGCACG *atpA*-R: CTTGAAACCGACAAGCCCAC	627	55°C
efp	*efp*-F: GGCAATTTGGATGAGCGAGCTC *efp*-R: CTTCACCTTTTCAAGATACTC	410	55°C
mutY	*mutY*-F: GTGGTTGTAGYTGGAAACTTTACAC *mutY*-R: CTTAAGCGTGTGTYTTTCTAGG	420	55°C
ppa	*ppa*-F: GGAGATTGCAATGAATTTAGA *ppa*-R: GTGGGGTTAARATCGTTAAATTG	398	55°C
trpC	*trpC*-F: TAGAATGCAAAAAAGCATCGCCCTC *trpC*-R: TAAGCCCGCACACTTTATTTTCGCC	456	55°C
ureI	*ureI*-F: AGGTTATTCGTAAGGTGCG *ureI*-R: GTTTAAATCCCTTAGATTGCC	585	55°C
yphC	*yphC*-F: CACGCCTATTTTTTTGACTAAAAAC *yphC*-R: CATTYACCCTCCCAATGATGC	510	55°C

atpA, encodes an ATP synthase alpha chain; *efp*, encodes an elongation factor P; *mutY*, encodes a DNA glycosylase; *ppa*, encodes an inorganic pyrophosphatase; *trpC*, encodes an anthranilate isomerase; *ureI*, encodes a urease subunit I; *yphC*: encodes a GTPase.

10.3389/fmicb.2023.1207878

TABLE 2 Dataset of alleles and STs of *Helicobacter pylori* in this study.

	Isolation source	ST	Allele						
			atpA	*efp*	*mutY*	*ppa*	*trpC*	*urel*	*yphC*
ATCC 43629		4,089	198	92	234	222	226	306	317
ATCC 700392		181	181	181	181	181	181	181	181
ATCC 51932		4,09	1,749	200	1,803	201	1,828	3,322	2,059
ATCC 700824		3,496	2,762	199	2,806	199	199	199	199
ATCC 43579		4,092	3,189	2,992	3,280	3,040	3,318	3,323	3,397
ATCC 49503		4,095	3,190	2,993	3,284	3,041	3,319	3,324	3,398
Patient 1	DB	3,689	2,886	2,697	2,945	2,774	2,996	2,992	3,063
Patient 2	DB	3,744	2,915	2,726	2,972	253	3,025	3,019	3,088
Patient 3	A	4,096	3,192	2,995	3,282	3,043	3,321	3,327	3,400
Patient 4	C	4,097	3,193	2,996	3,283	3,043	3,322	3,326	3,401
Patient 5	A	3,715	2,906	2,723	2,966	2,781	3,020	3,015	3,083
Patient 6	DB	3,696	2,895	2,697	2,945	2,774	2,996	2,992	3,063
Patient 7	A	3,710	2,903	2,699	2,947	2,768	2,998	2,994	3,065
	C	3,692	2,887	1,760	2,947	2,768	3,003	2,994	3,065
	DB	4,090	2,887	2,699	2,947	2,768	2,998	2,994	3,065
Patient 8	A	3,735	2,889	2,727	2,976	2,783	3,029	3,023	3,093
	C	3,735	2,889	2,727	2,976	2,783	3,029	3,023	3,093
	DB	3,735	2,889	2,727	2,976	2,783	3,029	3,023	3,093
Patient 9	C	3,731	2,888	2,700	2,960	2,770	2,999	2,995	3,066
	DB	3,731	2,888	2,700	2,960	2,770	2,999	2,995	3,066
Patient 10	A	3,734	2,917	1,892	2,973	448	3,026	3,020	3,090
	C	3,733	2,903	1,892	2,973	448	3,026	3,020	3,090
	DB	3,734	2,917	1,892	2,973	448	3,026	3,020	3,090
Patient 11	A	4,093	2,890	2,701	2,948	2,769	3,000	2,996	3,398
	C	4,094	3,191	2,994	3,281	3,042	3,320	3,325	3,071
Patient 12	C	3,693	2,891	2,702	2,949	2,391	3,001	1,069	3,063
	DB	3,693	2,891	2,702	2,949	2,391	3,001	1,069	3,063
Patient 13	A	3,699	1,987	1,870	2,961	2,415	3,008	3,002	3,071
	C	3,699	1,987	1,870	2,961	2,415	3,008	3,002	3,071
	DB	3,699	1,987	1,870	2,961	2,415	3,008	3,002	3,071
Patient 14	C	3,694	2,892	2,703	2,950	2,770	3,002	2,997	3,067
Patient 15	DB	3,736	2,919	2,728	2,977	2,784	3,030	3,024	3,094
Patient 16	A	3,695	2,894	2,705	2,951	6	3,004	2,998	3,068
	C	3,695	2,894	2,705	2,951	6	3,004	2,998	3,068
	DB	3,695	2,894	2,705	2,951	6	3,004	2,998	3,068
Patient 17	A	3,732	2,916	2,698	2,946	2,767	2,997	2,993	3,064
	C	3,709	2,908	2,698	2,946	2,767	2,997	2,993	3,064
Patient 18	A	3,697	2,903	2,704	2,952	2,771	3,006	3,000	3,069
	C	3,697	2,903	2,704	2,952	2,771	3,006	3,000	3,069
	DB	3,697	2,903	2,704	2,952	2,771	3,006	3,000	3,069
Patient 19	A	3,711	2,896	2,706	2,956	2,772	3,007	3,001	3,075
	C	3,698	2,896	2,706	2,956	2,772	3,007	3,001	3,070

(Continued)

TABLE 2 (Continued)

	Isolation source	ST	Allele						
			atpA	*efp*	*mutY*	*ppa*	*trpC*	*urel*	*yphC*
	DB	3,698	2,896	2,706	2,956	2,772	3,007	3,001	3,070
Patient 20	A	3,706	2,902	2,712	2,010	2,776	3,013	1,364	3,079
	C	3,706	2,902	2,712	2,010	2,776	3,013	1,364	3,079
	DB	3,747	2,902	2,712	2,978	2,776	3,013	1,364	3,079
Patient 21	A	3,700	2,897	2,707	2,953	448	3,009	3,003	3,072
		3,701	2,903	2,707	2,953	448	3,009	3,003	3,072
	DB	3,700	2,897	2,707	2,953	448	3,009	3,003	3,072
		3,701	2,903	2,707	2,953	448	3,009	3,003	3,072
		3,702	2,898	2,707	2,953	448	3,009	3,003	3,072
Patient 22	A	3,704	2,899	2,710	2,954	1,279	3,010	3,004	3,073
		3,746	2,903	2,710	2,954	1,279	3,010	3,004	3,073
	C	3,723	2,913	2,714	2,955	2,773	3,011	3,005	3,074
Patient 23	A	3,737	2,920	2,708	2,957	2,775	3,013	3,007	3,076
	C	3,737	2,920	2,708	2,957	2,775	3,013	3,007	3,076
	DB	3,737	2,920	2,708	2,957	2,775	3,013	3,007	3,076
Patient 24	A	3,705	2,901	2,711	2,958	932	3,014	3,008	3,077
	C	3,705	2,901	2,711	2,958	932	3,014	3,008	3,077
	DB	3,705	2,901	2,711	2,958	932	3,014	3,008	3,077
Patient 25	A	3,738	2,900	2,716	2,963	2,778	3,017	3,018	3,095
	C	3,738	2,900	2,716	2,963	2,778	3,017	3,018	3,095
	DB	3,738	2,900	2,716	2,963	2,778	3,017	3,018	3,095
Patient 26	A	3,739	2,904	2,721	2,964	2,779	3,018	3,017	3,081
	C	3,739	2,904	2,721	2,964	2,779	3,018	3,017	3,081
	DB	3,739	2,904	2,721	2,964	2,779	3,018	3,017	3,081
Patient 27	A	3,730	2,905	2,722	2,965	2,780	3,019	3,016	3,082
	C	3,730	2,905	2,722	2,965	2,780	3,019	3,016	3,082
Patient 28	C	3,740	2,893	2,709	2,962	502	3,005	3,010	3,080
	DB	3,740	2,893	2,709	2,962	502	3,005	3,010	3,080
Patient 29	A	3,708	2,903	2,713	2,959	2,777	3,015	3,009	3,078
	C	3,708	2,903	2,713	2,959	2,777	3,015	3,009	3,078
Patient 30	A	3,704	2,899	2,710	2,954	1,279	3,010	3,004	3,073
	C	3,723	2,913	2,714	2,955	2,773	3,011	3,005	3,074
Patient 31	A	3,716	2,907	2,715	2,967	s499	3,012	3,014	3,084
	C	3,716	2,907	2,715	2,967	499	3,012	3,014	3,084
		3,741	2,910	2,725	2,968	2,785	3,021	3,013	3,085
		3,718	2,911	2,724	2,969	2,782	3,022	3,012	3,086
Patient 32	A	3,748	2,909	2,717	2,982	56	3,023	3,011	3,087
Patient 33	A	3,720	963	2,718	2,970	7	3,016	2,999	3,089
		3,724	963	2,718	2,970	7	3,016	3,006	3,089
		3,743	2,912	2,719	2,970	1,466	3,024	3,006	3,089
		3,749	2,922	2,718	2,970	7	3,016	2,999	3,089
	C	3,720	963	2,718	2,970	7	3,016	2,999	3,089
		3,742	963	2,720	2,970	7	3,016	2,999	3,089

The newly emerged alleles and STs of the isolates from this study are indicated in red. A, antrum; C, corpus; DB, duodenal bulb.

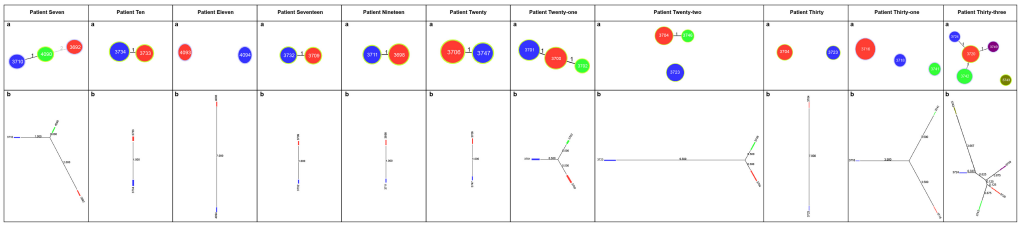


FIGURE 1
Evolutionary relationships among STs of *H. pylori* in patients with mixed infections. **(A)** Image showing the clonal relationships between the STs of *H. pylori*. Each circle represents an ST, and the size of the circle indicates the number of clinical isolates in the ST. The number on each line indicates the number of alleles with mutations. The unlinked STs imply that there are more than three different housekeeping genes between them. PHYLOViZ (goeBURST algorithm) was used to define the clonal relationships. **(B)** Image showing the evolutionary distance between the STs of *H. pylori*. Each square represents an ST; the larger the square, the more clinical isolates are included. Moreover, the longer the line and the larger the value on the line, the greater the genetic distance. PHYLOViZ (neighbor-joining algorithm) was used to define the evolutionary distance.

Results

H. pylori isolation

A total of 156 clinical isolates of *H. pylori* were obtained from 33 patients with dyspepsia (18 males and 15 females) at two hospitals: Ninghai First Hospital and the Affiliated Hospital of Ningbo University School of Medicine. Clinical data showed that among the 33 patients with gastric diseases investigated by gastric endoscopy, 6 (18%) had gastric ulcer, 4 (12%) had duodenal ulcer, and 2 (6%) had compound ulcer, whereas the remaining 13 (39%), 7 (21%), and 1 (3%) were diagnosed with chronic active gastritis, chronic superficial gastritis, and chronic atrophic gastritis by pathology, respectively. The clinical data of the 33 patients and information on the sampling of *H. pylori* isolates and the number of isolates are summarized in **Supplementary Table 1**.

Multilocus sequence typing results

Multilocus sequence typing (MLST) was used to classify 162 *H. pylori* strains into 55 sequence types (STs). The STs of the six *H. pylori* isolates in the clinical control group were ST3689, ST3744, ST4096, ST4097, ST3715, and ST3696. The STs of the six ATCC reference strains were ATCC 43629 for ST4089, ATCC 700392 for ST181, ATCC 51932 for ST4091, ATCC 700824 for ST3496, ATCC 43579 for ST4092, and ATCC 49503 for ST4095. MLST separated the six clinical control isolates and the six ATCC reference strains into twelve STs. This indicated that isolates from different patients possessed different STs and that MLST genotyping in this study was accurate. The STs of 156 *H. pylori* isolates from the 33 patients as well as 6 ATCC strains are shown in **Table 2**. The analysis of seven housekeeping genes in these strains revealed 246 new alleles (*atpA*, 39; *efp*, 37; *mutY*, 37; *ppa*, 23; *trpC*, 38; *ureI*, 37; *yphC*, 35) and 53 new STs.

Evolutionary relationships between isolates in patients with mixed infections

The different STs identified in each patient with mixed infections were used as genotyping data for the PHYLOViZ platform (Mendoza-Elizalde et al., 2019). The evolutionary relationships between isolates in the 11 patients with mixed infections are shown in **Figure 1**. The goeBURST algorithm was used at the TLV level to define clonal relationships. In the case of Patient 7, ST4090 was the major clonal complex with two linked STs (ST3710 and ST3692). The neighbor-joining algorithm was used to define the evolutionary distance; the longer the line, the greater the genetic distance. The number on the line represents how many housekeeping genes differ between these STs. For example, Patient 7 had one different housekeeping gene between ST3710 and ST4090, two differences between ST4090 and ST3692, and three differences between ST3710 and ST3692.

According to the evolutionary relationships and the polymorphisms of housekeeping genes of these isolates (**Table 3**), four of the eleven patients with mixed infections had isolates with seven different housekeeping genes, among which three patients (Patients 11, 22, and 30) showed different STs colonized at different sites (antrum, corpus), and one patient (Patient 31) exhibited three different STs at the same site (corpus). Moreover, two patients (Patients 7 and 33) had isolates with up to three and five different housekeeping genes, respectively. Eight of the eleven patients had isolates with only one housekeeping gene change, with *atpA* accounting for 75% (6/8), *mutY* for 12.5% (1/8) and *yphC* for 12.5% (1/8). Based on clonal relationships at the TLV level, related strains had the same STs or STs that differed by only very few alleles, whereas unrelated strains had STs that differed by more than three alleles. Therefore, the isolates from five patients (Patients 11, 22, 30, 31 and 33) were considered unrelated strains, and even isolates from four of these patients were considered completely unrelated strains (strains with all seven housekeeping genes different).

Phylogenetic analysis (**Figure 2**) showed that among the patients with mixed infections, the strains from four patients (Patients 11, 22, 30, and 31) were genetically distant and did not cluster together, indicating no evolutionary relatedness. The seven housekeeping genes of the isolates from these four patients were all different.

Phylogeography

The phylogenetic tree revealed that the majority (153) of the 156 *H. pylori* clinical isolates belonged to the hspEAsia subgroup,

10.3389/fmicb.2023.1207878

TABLE 3 Polymorphisms of housekeeping genes of *Helicobacter pylori* isolates obtained from 11 patients with mixed infections.

	Gene	Polymorphic sites (S)	Number of haplotypes (h)	Haplotype diversity (Hd)	Nucleotide diversity (Pi)
Patient 7	*atpA*	6	2	0.667	0.00638
	efp	4	2	0.667	0.00650
	mutY	0	1	0.000	0.00000
	ppa	0	1	0.000	0.00000
	trpC	1	2	0.667	0.00146
	ureI	0	1	0.000	0.00000
	yphC	0	1	0.000	0.00000
	Concatenated	11	3	1.000	0.00215
Patient 10	*atpA*	9	2	0.667	0.00957
	efp	0	1	0.000	0.00000
	mutY	0	1	0.000	0.00000
	ppa	0	1	0.000	0.00000
	trpC	0	1	0.000	0.00000
	ureI	0	1	0.000	0.00000
	yphC	0	1	0.000	0.00000
	Concatenated	9	2	0.667	0.00176
Patient 11	*atpA*	16	2	1.000	0.02552
	efp	11	2	1.000	0.02683
	mutY	17	2	1.000	0.04048
	ppa	6	2	1.000	0.01508
	trpC	116	2	1.000	0.25439
	ureI	6	2	1.000	0.01026
	yphC	19	2	1.000	0.03725
	Concatenated	191	2	1.000	0.05608
Patient 17	*atpA*	3	2	1.000	0.00478
	efp	0	1	0.000	0.00000
	mutY	0	1	0.000	0.00000
	ppa	0	1	0.000	0.00000
	trpC	0	1	0.000	0.00000
	ureI	0	1	0.000	0.00000
	yphC	0	1	0.000	0.00000
	Concatenated	3	2	1.000	0.00088
Patient 19	*atpA*	0	1	0.000	0.00000
	efp	0	1	0.000	0.00000
	mutY	0	1	0.000	0.00000
	ppa	0	1	0.000	0.00000
	trpC	0	1	0.000	0.00000
	ureI	0	1	0.000	0.00000
	yphC	15	2	0.667	0.01961
	Concatenated	15	2	0.667	0.00294
Patient 20	*atpA*	0	1	0.000	0.00000
	efp	0	1	0.000	0.00000
	mutY	1	2	0.667	0.00159
	ppa	0	1	0.000	0.00000

(Continued)

10.3389/fmicb.2023.1207878

TABLE 3 (Continued)

	Gene	Polymorphic sites (S)	Number of haplotypes (h)	Haplotype diversity (Hd)	Nucleotide diversity (Pi)
	trpC	0	1	0.000	0.00000
	ureI	0	1	0.000	0.00000
	yphC	0	1	0.000	0.00000
	Concatenated	1	2	0.667	0.00020
Patient 21	*atpA*	8	3	0.581	0.00673
	efp	0	1	0.000	0.00000
	mutY	0	1	0.000	0.00000
	ppa	0	1	0.000	0.00000
	trpC	0	1	0.000	0.00000
	ureI	0	1	0.000	0.00000
	yphC	0	1	0.000	0.00000
	Concatenated	8	3	0.581	0.00124
Patient 22	*atpA*	13	3	0.600	0.00787
	efp	14	2	0.533	0.01821
	mutY	12	2	0.533	0.01524
	ppa	7	2	0.533	0.00938
	trpC	14	2	0.533	0.01637
	ureI	15	2	0.533	0.01368
	yphC	13	2	0.533	0.01359
	Concatenated	88	3	0.600	0.01319
Patient 30	*atpA*	8	2	0.485	0.00619
	efp	14	2	0.485	0.01656
	mutY	12	2	0.485	0.01385
	ppa	7	2	0.485	0.00853
	trpC	14	2	0.485	0.01489
	ureI	15	2	0.485	0.01243
	yphC	13	2	0.485	0.01236
	Concatenated	83	2	0.485	0.01182
Patient 31	*atpA*	15	3	0.242	0.00368
	efp	11	3	0.242	0.00467
	mutY	16	3	0.242	0.00760
	ppa	7	3	0.242	0.00247
	trpC	19	3	0.242	0.00592
	ureI	8	3	0.242	0.00208
	yphC	16	3	0.242	0.00606
	Concatenated	92	3	0.242	0.00452
Patient 33	*atpA*	4	3	0.205	0.00082
	efp	3	3	0.526	0.00163
	mutY	0	1	0.000	0.00000
	ppa	1	2	0.105	0.00026
	trpC	5	2	0.105	0.00115
	ureI	1	2	0.199	0.00034
	yphC	0	1	0.000	0.00000
	Concatenated	14	5	0.649	0.00059

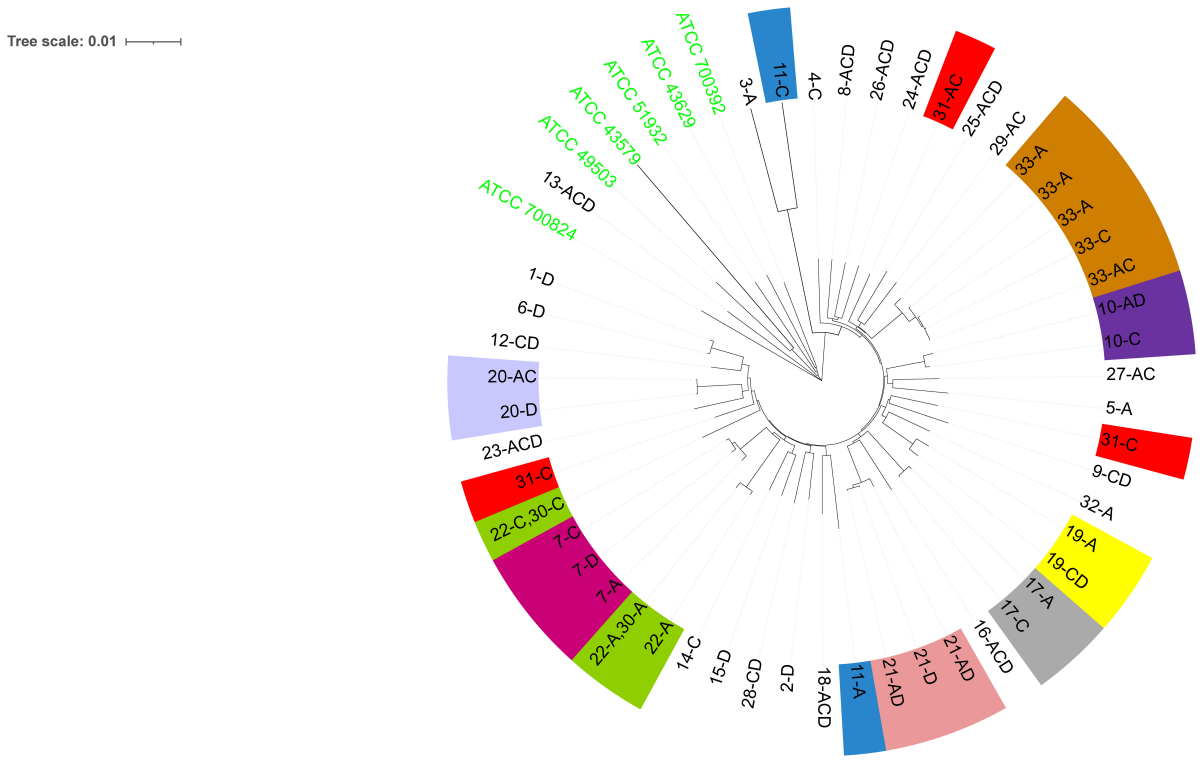


FIGURE 2
The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei, 1987). The optimal tree is shown. The evolutionary distances were computed using the Kimura (1980) 2-parameter method and were in the units of the number of base substitutions per site. This analysis involved 55 nucleotide sequences. All positions containing gaps and missing data were eliminated (complete deletion option). A total of 3,406 positions were included in the final dataset. Evolutionary analyses were conducted using MEGA11 (Tamura et al., 2021). The six ATCC strains shown are the reference strains. Numbers 1–33 are patient numbers. Strains from the same patient are indicated by the same color. A, antrum; C, corpus; D, duodenum bulb.

whereas the three isolates (ST3699) from Patient 13 belonged to the hpEurope group (**Figure 3**). Among the reference strains, ATCC 700824 belonged to hspWAfrica, and the rest belonged to hpEurope.

Discussion

Helicobacter pylori is the most common infectious pathogen causing gastrointestinal diseases in humans, and can cause chronic gastritis, peptic ulcer, and gastric cancer (Blaser and Atherton, 2004; Burkitt et al., 2017; Camilo et al., 2017). Individuals can be infected by one or more types of *H. pylori*, which is known as multiple infections (Didelot et al., 2013; Palau et al., 2020), co-infection (Didelot et al., 2013; Seo et al., 2019; Mi et al., 2021), or mixed infections (Kao et al., 2014; Lai et al., 2016). Ben Mansour et al. (2016) distinguished the concepts of multiple infections and mixed infections, and defined multiple infections as an individual with two or more genetically different *H. pylori* isolates, whereas they defined mixed infections as an individual with two or more *H. pylori* isolates with different antibiotic sensitivity characteristics. It would contradict much of the literature to define mixed infections as simply different strains involving differences in antibiotic susceptibility characteristics. In essence, mixed infections are the colonization of the upper gastrointestinal mucosa of an individual by strains with heterogeneous biological characteristics, which can be reflected in the phenotypic characteristics of the strains, such as virulence factors, heteroresistance (Lai et al., 2016; Mi et al., 2021), and as well as strains with heterogeneous DNA.

The diagnosis of *H. pylori* mixed infection is mainly through genotyping techniques, including detection of virulence genes (*cagA*, *vacA*, *iceA*), random amplified polymorphic DNA (RAPD), MLST, and whole-genome sequencing (WGS). WGS is the gold standard for research, but its application is limited due to its high cost and complexity of analysis (Wang and Gu, 2018). MLST is the standard method of molecular typing of bacteria because of its high resolution and repeatability (Raaf et al., 2017). In addition, it can be used to determine the genotype of ancestral strains in individual patients (Mendoza-Elizalde et al., 2019). Although RAPD is low cost and rapid, the PCR-based RAPD fingerprinting pattern has low reproducibility across different experiments (Yokota et al., 2015). Moreover, RAPD has limitations in determining *H. pylori* mixed infections: researchers can only subjectively determine whether there is relatedness between different strains by DNA fingerprinting and cannot determine the evolutionary relationships of these isolates from a single host. It is unclear whether strain diversity is a result of multiple-strain infections or microevolution occurring in a single ancestral strain. In this study, we used MLST to explore the occurrence

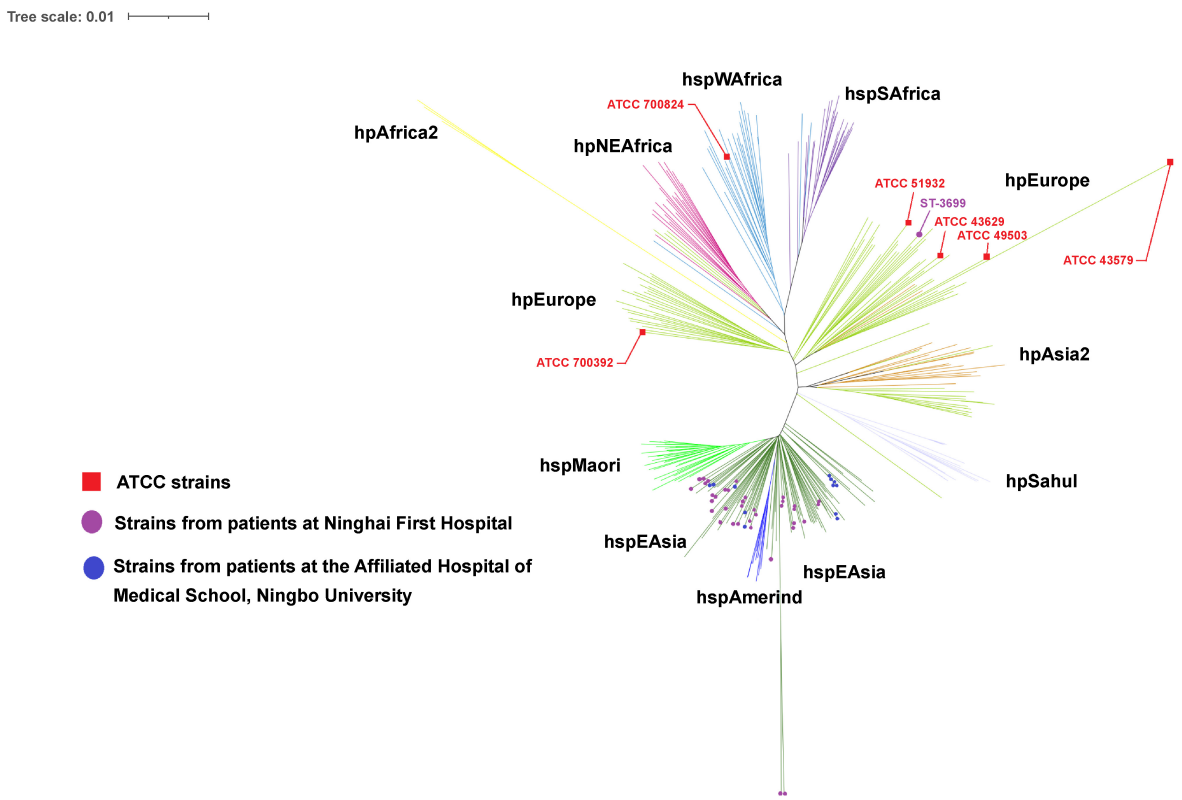


FIGURE 3
Phylogeography of the analyzed strains. The phylogeography was inferred using the neighbor-joining method (Saitou and Nei, 1987). The optimal tree is shown. The evolutionary distances were computed using the Kimura (1980) 2-parameter method and were in the units of the number of base substitutions per site. This analysis involved 55 (in this study) + 284 [from the *Helicobacter pylori* PubMLST database (http://pubmlst.org/organisms/helicobacter-pylori)] nucleotide sequences. All positions containing gaps and missing data were eliminated (complete deletion option). A total of 3,406 positions were included in the final dataset. Phylogeographical analyses were conducted using MEGA11 (Tamura et al., 2021).

of strain diversity. We concatenated seven housekeeping genes of each *H. pylori* strain and used the PHYLOViZ and neighbor-joining algorithm to show the evolutionary relationships among the strains. We found that most of the colonized strains of patients with mixed infections differed by only one housekeeping gene, suggesting that different genotypes may originate from a single ancestral strain. The evolutionary relatedness of strains can be identified by comparing STs (Mendoza-Elizalde et al., 2019), and in this study, five patients had more than three different housekeeping genes between their respective STs, which indicated that they were infected with unrelated strains. Moreover, four of these patients had isolates with seven different housekeeping genes, indicating infection with completely unrelated strains. The prevalence of *H. pylori* mixed infections involving completely unrelated strains was non-negligible. Notably, we found that unrelated *H. pylori* strains could colonize the same gastric site (corpus) or different sites (antrum and corpus), indicating that both the gastric antrum and corpus can be colonized by mixed unrelated strains, and that the same site can be infected with multiple different types of *H. pylori*. Previous studies have revealed the existence of a microevolutionary mechanism through slightly different DNA fingerprinting patterns among *H. pylori* isolates from a single host (Wong et al., 2001; Kim et al., 2003; Carroll et al., 2004; Farzi et al., 2019), and our study further revealed a mechanism of mixed infections based on MLST technology.

The prevalence of *H. pylori* mixed infections varies widely according to region. For example, the prevalence of mixed infections in developed countries/regions is generally lower than that in developing countries/regions (Ben Mansour et al., 2016). In this study, we used MLST and determined that the prevalence of *H. pylori* mixed infections among patients in Ningbo, China, was 41%, which is higher than that reported in other regions of China, such as Hong Kong (24% by RAPD) (Wong et al., 2001) and Taiwan (28.6% by detection of virulence genes) (Lai et al., 2016), but lower than that in Guiyang (76.9% by RAPD) (Mi et al., 2021). It remains to be investigated whether the differences in prevalence are related to the measurement method, other than the geographic region. This study confirmed that MLST is more reliable than RAPD for determining the prevalence of mixed infections. In this study, 12 different STs were obtained by MLST from *H. pylori* isolates isolated from 6 unrelated patient control groups and 6 ATCC controls, whereas only two different profiles from the clinical isolates were obtained by RAPD (data not shown), indicating that four isolates were not distinguished as different based on the RAPD type. However, almost all studies on mixed infections have not included controls and reference strains for *H. pylori* isolates from patients and ATCC strains (Wong et al., 2001; Kim et al., 2003;

Carroll et al., 2004; Kao et al., 2014; Ben Mansour et al., 2016; Lai et al., 2016; Seo et al., 2019; Palau et al., 2020; Mi et al., 2021).

It is unclear whether the prevalence of *H. pylori* mixed infections may varies according to the presence of disease, and sex. The prevalence of mixed infections in duodenal ulcer patients was higher than other diseases (Lai et al., 2016), but other studies did not support this conclusion (Kim et al., 2004; Ben Mansour et al., 2016). One report suggested that the prevalence of *H. pylori* mixed infections was significantly higher in women than in men (Kibria et al., 2015), but other reports confirmed that the prevalence was unrelated to sex (Ben Mansour et al., 2016). The data from the **Supplementary Table 1** showed that, in this study, the prevalence of *H. pylori* mixed infection was independent of disease status or sex. Whether these inconsistent findings are caused by different assay methods needs to be confirmed.

Reasons for the variation in the prevalence of mixed infections may also include the effects of the sampling strategies, such as biopsy specimens from one vs. several gastrointestinal sites and one vs. multiple colonies from each biopsy. The assay using *H. pylori* only from the antrum may not be representative of *H. pylori* populations in the entire stomach, and Wong et al. (2001) suggested that several different sites need to be assayed. In this study, we used *H. pylori* isolates from the gastric antrum, corpus, and/or duodenum bulb for the assays.

In this study, we identified 246 new alleles and 53 new STs associated with *H. pylori* infection. *H. pylori* is a highly genetically diverse species (Suerbaum et al., 1998). Analysis of the polymorphisms of housekeeping genes of *H. pylori* obtained from patients with mixed infections showed higher Hd values, which indicated that haplotypes forming the populations in each patient were very divergent, thus confirming the high diversity of *H. pylori*. The number of polymorphic sites revealed that *H. pylori* can mutate, making *H. pylori* more adaptable to specific ecological niches in individual hosts (Kang and Blaser, 2006; Naito et al., 2006).

Phylogenetic analyses of the seven concatenated housekeeping genes revealed different geographical groups of *H. pylori*. In this study, *H. pylori* was most prominently represented by the hspEAsia subgroup, and for the first time, three isolates (ST3699) recovered from a local patient with no history of living abroad were found to belong to the hpEurope group. MLST is a useful tool for tracking human migration (Yamaoka, 2009), enabling us to infer human migration routes or colonization history based on the geographic grouping of *H. pylori*. *H. pylori* strains that colonize in patients from Ningbo belong predominantly to the hspEAsia subgroup, which is consistent with the geographical location of Ningbo (eastern Asia).

In this study, we used a self-developed kit, Gu's kit, for rapid isolation of *H. pylori*, which is a new technique for obtaining isolates rapidly and efficiently. Studies have reported that some genotypes were lost during culture passages (Ren et al., 2012) and eventually failed to obtain isolates (Arévalo-Jaimes et al., 2019), which is detrimental to the study of mixed infections. This new and rapid culture technique can effectively shorten culture time and reduce the loss of genetic representation during subculture passages, which can truly reflect the patient's *H. pylori* mixed infections.

The main limitation of this study is the small sample size, which did not enable us to analyze the effect of the specific isolation scheme on the prevalence of mixed infections detected. Therefore, these results need to be confirmed by larger studies.

In conclusion, the prevalence of *H. pylori* mixed infections in patients from Ningbo, China, is high. Moreover, the prevalence of *H. pylori* mixed infections involving unrelated strains was 46%, and the prevalence of *H. pylori* mixed infections involving completely unrelated strains was 36%, which may be related to the use of new *H. pylori* isolation and culture techniques. In this study, 246 new alleles and 53 new STs were discovered for the first time. Most *H. pylori* strains in Ningbo belong to the hspEAsia subgroup, whereas very few strains belong to the hpEurope group.

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 this article/**Supplementary material**.

Ethics statement

The studies involving humans were approved by the Medical Research Committee of the Ningbo University School of Medicine. 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

HG conceived the study, collected gastrointestinal biopsy samples, and reviewed the manuscript. YZ performed the experiment, generated the sequence data, performed bioinformatics analysis, and prepared the manuscript. ZS, WC, WY, and CZ collected gastrointestinal biopsy samples. AL participated in the analysis of the results. MZ generated the sequence data. HY performed some of the experiments. All authors read and approved the final manuscript.

Funding

This study was supported by the Zhejiang Provincial Natural Science Foundation (LZ14H200001). The funder had no role in the design of the study or collection, analysis, and interpretation of data or in writing the manuscript.

Acknowledgments

We would like to thank Editage (www.editage.cn) for English language editing.

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.2023.1207878/full#supplementary-material

References

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z. J., et al. (1999). Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol Microbiol.* 32, 459–470. doi: 10.1046/j.1365-2958.1999.01382.x

Arévalo-Jaimes, B. V., Rojas-Rengifo, D. F., Jaramillo, C. A., de Molano, B. M., Vera-Chamorro, J. F., and Del Pilar Delgado, M. (2019). Genotypic determination of resistance and heteroresistance to clarithromycin in *Helicobacter pylori* isolates from antrum and corpus of Colombian symptomatic patients. *BMC Infect. Dis.* 19:546. doi: 10.1186/s12879-019-4178-x

Ben Mansour, K., Fendri, C., Battikh, H., Garnier, M., Zribi, M., Jlizi, A., et al. (2016). Multiple and mixed *Helicobacter pylori* infections: Comparison of two epidemiological situations in Tunisia and France. *Infect. Genet. Evol.* 37, 43–48. doi: 10.1016/j.meegid.2015.10.028

Blaser, M. J., and Atherton, J. C. (2004). *Helicobacter pylori* persistence: Biology and disease. *J. Clin. Invest.* 113, 321–333. doi: 10.1172/JCI20925

Blaser, M. J., and Berg, D. E. (2001). *Helicobacter pylori* genetic diversity and risk of human disease. *J. Clin. Invest.* 107, 767–773. doi: 10.1172/JCI12672

Burkitt, M. D., Duckworth, C. A., Williams, J. M., and Pritchard, D. M. (2017). *Helicobacter pylori*-induced gastric pathology: Insights from *in vivo* and *ex vivo* models. *Dis. Model. Mech.* 10, 89–104. doi: 10.1242/dmm.027649

Camilo, V., Sugiyama, T., and Touati, E. (2017). Pathogenesis of *Helicobacter pylori* infection. *Helicobacter* 22, 24–31. doi: 10.1111/hel.12405

Carroll, I. M., Ahmed, N., Beesley, S. M., Khan, A. A., Ghousunnissa, S., Moráin, C. A. Ó, et al. (2004). Microevolution between paired antral and paired antrum and corpus *Helicobacter pylori* isolates recovered from individual patients. *J. Med. Microbiol.* 53, 669–677. doi: 10.1099/jmm.0.05440-0

Didelot, X., Nell, S., Yang, I., Woltemate, S., van der Merwe, S., and Suerbaum, S. (2013). Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc. Natl. Acad. Sci. U. S. A.* 110, 13880–13885. doi: 10.1073/pnas.1304681110

Farzi, N., Behzad, C., Hasani, Z., Alebouyeh, M., Zojaji, H., and Zali, M. R. (2019). Characterization of clarithromycin heteroresistance among *Helicobacter pylori* strains isolated from the antrum and corpus of the stomach. *Folia Microbiol.* 64, 143–151. doi: 10.1007/s12223-018-0637-9

Farzi, N., Malekian, T., Alebouyeh, M., Vaziri, F., and Zali, M. R. (2015). Genotype diversity and quasispecies development of *Helicobacter pylori* in a single host. *Jpn. J. Infect. Dis.* 68, 176–180. doi: 10.7883/yoken.JJID.2014.165

Gutiérrez-Escobar, A. J., Trujillo, E., Acevedo, O., and Bravo, M. M. (2017). Phylogenomics of Colombian *Helicobacter pylori* isolates. *Gut Pathog.* 9:52. doi: 10.1186/s13099-017-0201-1

Kang, J., and Blaser, M. J. (2006). Bacterial populations as perfect gases: Genomic integrity and diversification tensions in *Helicobacter pylori*. *Nat. Rev. Microbiol.* 4, 826–836. doi: 10.1038/nrmicro1528

Kao, C. Y., Lee, A. Y., Huang, A. H., Song, P. Y., Yang, Y. J., Sheu, S. M., et al. (2014). Heteroresistance of *Helicobacter pylori* from the same patient prior to antibiotic treatment. *Infect. Genet. Evol.* 23, 196–202. doi: 10.1016/j.meegid.2014.02.009

Kersulyte, D., Mukhopadhyay, A. K., Velapatiño, B., Su, W., Pan, Z., Garcia, C., et al. (2000). Differences in genotypes of *Helicobacter pylori* from different human populations. *J. Bacteriol.* 182, 3210–3218. doi: 10.1128/JB.182.11.3210-3218.2000

Kibria, K. M., Hossain, M. E., Sultana, J., Sarker, S. A., Bardhan, P. K., Rahman, M., et al. (2015). The prevalence of mixed *Helicobacter pylori* infections in symptomatic and asymptomatic subjects in Dhaka, Bangladesh. *Helicobacter* 20, 397–404. doi: 10.1111/hel.12213

Kim, J. J., Kim, J. G., and Kwon, D. H. (2003). Mixed-infection of antibiotic susceptible and resistant *Helicobacter pylori* isolates in a single patient and underestimation of antimicrobial susceptibility testing. *Helicobacter* 8, 202–206. doi: 10.1046/j.1523-5378.2003.00145.x

Kim, J. W., Kim, J. G., Chae, S. L., Cha, Y. J., and Park, S. M. (2004). High prevalence of multiple strain colonization of *Helicobacter pylori* in Korean patients: DNA diversity among clinical isolates from the gastric corpus, antrum and duodenum. *Korean. J. Intern. Med.* 19, 1–9. doi: 10.3904/kjim.2004.19.1.1

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kusters, J. G., van Vliet, A. H., and Kuipers, E. J. (2006). Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol. Rev.* 19, 449–490. doi: 10.1128/CMR.00054-05

Lai, C. H., Huang, J. C., Chiang-Ni, C., Li, J. P., Wu, L. T., Wu, H. S., et al. (2016). Mixed infections of *Helicobacter pylori* isolated from patients with gastrointestinal diseases in Taiwan. *Gastroenterol. Res. Pract.* 2016:7521913. doi: 10.1155/2016/7521913

Marshall, D. G., Dundon, W. G., Beesley, S. M., and Smyth, C. J. (1998). *Helicobacter pylori* – A conundrum of genetic diversity. *Microbiology* 144, 2925–2939. doi: 10.1099/00221287-144-11-2925

Matteo, M. J., Granados, G., Pérez, C. V., Olmos, M., Sanchez, C., and Catalano, M. (2007). *Helicobacter pylori* cag pathogenicity island genotype diversity within the gastric niche of a single host. *J. Med. Microbiol.* 56, 664–669. doi: 10.1099/jmm.0.46885-0

Mégraud, F., Lehours, P., and Vale, F. F. (2016). The History of *Helicobacter pylori*: From phylogeography to paleomicrobiology. *Clin. Microbiol. Infect.* 22, 922–927. doi: 10.1016/j.cmi.2016.07.013

Mendoza-Elizalde, S., Cortés-Márquez, A. C., Zuñiga, G., Cerritos, R., Valencia-Mayoral, P., Sánchez, A. C., et al. (2019). Inference from the analysis of genetic structure of *Helicobacter pylori* strains isolates from two paediatric patients with recurrent infection. *BMC Microbiol.* 19:184. doi: 10.1186/s12866-019-1554-z

Mi, M., Wu, F., Zhu, J., Liu, F., Cui, G., Wen, X., et al. (2021). Heterogeneity of *Helicobacter pylori* strains isolated from patients with gastric disorders in Guiyang, China. *Infect. Drug Resist.* 14, 535–545. doi: 10.2147/IDR.S287631

Naito, M., Yamazaki, T., Tsutsumi, R., Higashi, H., Onoe, K., Yamazaki, S., et al. (2006). Influence of EPIYA-repeat polymorphism on the phosphorylation-dependent biological activity of *Helicobacter pylori* CagA. *Gastroenterology* 130, 1181–1190. doi: 10.1053/j.gastro.2005.12.038

Osaki, T., Konno, M., Yonezawa, H., Hojo, F., Zaman, C., Takahashi, M., et al. (2015). Analysis of intra-familial transmission of *Helicobacter pylori* in Japanese families. *J. Med. Microbiol.* 64, 67–73. doi: 10.1099/jmm.0.080507-0

Palau, M., Piqué, N., Comeau, A. M., Douglas, G. M., Ramírez-Lázaro, M. J., Lario, S., et al. (2020). Detection of *Helicobacter pylori* microevolution and multiple infection from gastric biopsies by housekeeping gene amplicon sequencing. *Pathogens* 9:97. doi: 10.3390/pathogens9020097

Patra, R., Chattopadhyay, S., De, R., Ghosh, P., Ganguly, M., Chowdhury, A., et al. (2012). Multiple infection and microdiversity among *Helicobacter pylori* isolates in a single host in India. *PLoS One* 7:e43370. doi: 10.1371/journal.pone.0043370

Raaf, N., Amhis, W., Saoula, H., Abid, A., Nakmouche, M., Balamane, A., et al. (2017). Prevalence, antibiotic resistance, and MLST typing of *Helicobacter pylori* in Algiers, Algeria. *Helicobacter* 22, e12446. doi: 10.1111/hel.12446

Ren, L., Liao, Y. L., Song, Y., Guo, Y., Mao, X. H., Xie, Q. H., et al. (2012). High frequency variations of *Helicobacter pylori* isolates in individual hosts in a Chinese population. *Int. J. Infect. Dis.* 16, e358–e363. doi: 10.1016/j.ijid.2012.01.006

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Salama, N. R., Gonzalez-Valencia, G., Deatherage, B., Aviles-Jimenez, F., Atherton, J. C., Graham, D. Y., et al. (2007). Genetic analysis of *Helicobacter pylori* strain populations colonizing the stomach at different times postinfection. *J. Bacteriol.* 189, 3834–3845. doi: 10.1128/JB.01696-06

Seo, J. W., Park, J. Y., Shin, T. S., and Kim, J. G. (2019). The analysis of virulence factors and antibiotic resistance between *Helicobacter pylori* strains isolated from gastric antrum and body. *BMC Gastroenterol.* 19:140. doi: 10.1186/s12876-019-1062-5

Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., et al. (1998). Free recombination within *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U. S. A.* 95, 12619–12624. doi: 10.1073/pnas.95.21.12619

Talebi Bezmin Abadi, A., and Perez-Perez, G. (2016). Role of *dupA* in Virulence of *Helicobacter pylori*. *World J. Gastroenterol.* 22, 10118–10123. doi: 10.3748/wjg.v22.i46.10118

Tamura, K., Stecher, G., and Mega, K. S. (2021). MEGA11: Molecular evolutionary genetics analysis, version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Vazirzadeh, J., Karbasizadeh, V., Falahi, J., Moghim, S., Narimani, T., and Rafiei, R. (2022). Genetic diversity of *Helicobacter pylori* isolates from patients with gastric diseases in Isfahan. *Adv. Biomed. Res.* 11:4. doi: 10.4103/abr.abr_25_21

Wang, J., and Gu, H. (2018). Research progress on genotyping of *Helicobacter pylori*. *J. Zhejiang Univ. Med. Sci.* 47, 97–103. doi: 10.3785/j.issn.1008-9292.2018.02.14

Wong, B. C., Wang, W. H., Berg, D. E., Fung, F. M., Wong, K. W., Wong, W. M., et al. (2001). High prevalence of mixed infections by *Helicobacter pylori* in Hong Kong: Metronidazole sensitivity and overall genotype. *Aliment. Pharmacol. Ther.* 15, 493–503. doi: 10.1046/j.1365-2036.2001.00949.x

Yamaoka, Y. (2009). *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infect.* 15, 829–834. doi: 10.1111/j.1469-0691.2009.02967.x

Yokota, S., Konno, M., Fujiwara, S., Toita, N., Takahashi, M., Yamamoto, S., et al. (2015). Intrafamilial, preferentially mother-to-child and intraspousal, *Helicobacter pylori* infection in Japan determined by mutilocus sequence typing and random amplified polymorphic DNA fingerprinting. *Helicobacter* 20, 334–342. doi: 10.1111/hel.12217

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 August 2023
DOI 10.3389/fgene.2023.1225248

Check for updates

OPEN ACCESS

EDITED BY
James Hane,
Curtin University, Australia

REVIEWED BY
Alexander Schönhuth,
Centrum Wiskunde & Informatica, Netherlands
Paola Bonizzoni,
University of Milano-Bicocca, Italy

*CORRESPONDENCE
Joep de Ligt,
joep.deligt@esr.cri.nz

RECEIVED 19 May 2023
ACCEPTED 01 August 2023
PUBLISHED 10 August 2023

CITATION
Yang Z, Guarracino A, Biggs PJ, Black MA, Ismail N, Wold JR, Merriman TR, Prins P, Garrison E and de Ligt J (2023), Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of *Neisseria meningitidis* leveraging Oxford Nanopore long reads. *Front. Genet.* 14:1225248. doi: 10.3389/fgene.2023.1225248

COPYRIGHT
© 2023 Yang, Guarracino, Biggs, Black, Ismail, Wold, Merriman, Prins, Garrison and de Ligt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of *Neisseria meningitidis* leveraging Oxford Nanopore long reads

Zuyu Yang[1], Andrea Guarracino[2,3], Patrick J. Biggs[4,5], Michael A. Black[6], Nuzla Ismail[6], Jana Renee Wold[7], Tony R. Merriman[6,8], Pjotr Prins[2], Erik Garrison[2] and Joep de Ligt[1]*

[1]Institute of Environmental Science and Research, Porirua, New Zealand, [2]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, United States, [3]Genomics Research Centre, Human Technopole, Milan, Italy, [4]Molecular Biosciences Group, School of Natural Sciences, Massey University, Palmerston North, New Zealand, [5]Molecular Epidemiology and Public Health Laboratory, School of Veterinary Science, Massey University, Palmerston North, New Zealand, [6]Department of Biochemistry, University of Otago, Dunedin, New Zealand, [7]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [8]Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, United States

Whole genome sequencing has revolutionized infectious disease surveillance for tracking and monitoring the spread and evolution of pathogens. However, using a linear reference genome for genomic analyses may introduce biases, especially when studies are conducted on highly variable bacterial genomes of the same species. Pangenome graphs provide an efficient model for representing and analyzing multiple genomes and their variants as a graph structure that includes all types of variations. In this study, we present a practical bioinformatics pipeline that employs the PanGenome Graph Builder and the Variation Graph toolkit to build pangenomes from assembled genomes, align whole genome sequencing data and call variants against a graph reference. The pangenome graph enables the identification of structural variants, rearrangements, and small variants (e.g., single nucleotide polymorphisms and insertions/deletions) simultaneously. We demonstrate that using a pangenome graph, instead of a single linear reference genome, improves mapping rates and variant calling for both simulated and real datasets of the pathogen *Neisseria meningitidis*. Overall, pangenome graphs offer a promising approach for comparative genomics and comprehensive genetic variation analysis in infectious disease. Moreover, this innovative pipeline, leveraging pangenome graphs, can bridge variant analysis, genome assembly, population genetics, and evolutionary biology, expanding the reach of genomic understanding and applications.

KEYWORDS

pangenome graphs, infectious diseases, genomic surveillance, comparative genomics, genetic variation, long-read sequencing, genome assembly

Introduction

Over the last two decades, whole genome sequencing (WGS) has become an indispensable tool in infectious disease research, surveillance, and control (Didelot et al., 2012; Gardy and Loman, 2018). Rapid advancements in sequencing technologies and bioinformatic analysis have facilitated the generation of high-quality genomic data at an unprecedented scale (Goodwin et al., 2016; Nurk et al., 2022). WGS has enabled researchers to track and monitor the spread and evolution of pathogens, investigate outbreaks, identify drug resistance markers, and develop diagnostic assays and vaccines (Koser et al., 2014; Walker et al., 2015; Quick et al., 2016; Chen et al., 2021; Yang et al., 2021). Its utility has been especially evident in the SARs-CoV-2 pandemic, enabling real-time tracking of the pandemic (Geoghegan et al., 2021) and identification of transmission chains (Geoghegan et al., 2020). Additionally, WGS has provided valuable insights into the genetic diversity, population structure, and functional characteristics of various pathogens, thereby shaping our understanding of the molecular mechanisms driving their virulence and transmission (Holt et al., 2015).

Currently, genomic surveillance concentrates on monitoring lineages and establishing transmission links between cases. Analysis is mainly dependent on mutations in the core genome (the genomic regions that are common to all isolates being analyzed at that time), using one linear genome as a reference. Bacteria genomes are highly variable, with genomic rearrangements and different-scale deletion or insertion events being common (Darmon and Leach, 2014). Using a single reference approach, variations in the accessory genome (regions not shared by all the genomes) are not detected, suggesting we may miss important variations and introduce biases due to the selection of the reference genome. Consequently, the alignment of sequencing data against a single reference genome may lead to inaccurate or incomplete variant identification (Garrison and Math, 2012). Moreover, the linear representation of a genome fails to capture the complexity of genomic rearrangements, duplications, and structural variants (SVs) that are critical for understanding pathogen evolution and adaptation, especially in highly recombinogenic species (Eizenga et al., 2020; Colquhoun et al., 2021). Viruses, responsible for many infectious diseases, possess highly variable genomes that complicate genomic surveillance (Sanjuán and Domingo-Calap, 2016). These tiny pathogens can rapidly evolve and adapt to changing environments, with the potential to jump species barriers, as seen with the emergence of SARS, MERS, and COVID-19 (Xu et al., 2004; Mohd et al., 2016; Plowright et al., 2017; Lu et al., 2020). Viral genomes, particularly those of RNA viruses, are characterized by high mutation rates which can lead to the emergence of new viral strains with altered virulence or transmissibility (Domingo et al., 2021). Accounting for the variability and unique characteristics of viral genomes is essential for comprehensive disease monitoring and management.

To overcome these limitations, pangenome graphs have emerged as an alternative approach for representing and analyzing multiple genomes and their variants (Garrison et al., 2018; Rakocevic et al., 2019; Liao et al., 2023; The Computational Pan-Genomics Consortium, 2018). A pangenome graph is a graph-based data structure that captures the entire genomic diversity of a set of related genomes by incorporating all types of variation, including SVs, rearrangements, and small variants (e.g., single nucleotide polymorphisms (SNPs) and insertions/deletions) (Paten et al., 2017; Marschall et al., 2018). By representing collections of genomes and their alignments as graphs, pangenome graphs allow for more accurate and comprehensive genetic variation analysis, as they provide a unified framework to compare and analyze diverse genomes, overcoming the biases associated with single linear reference genomes (Paten et al., 2017).

Different methods are available for constructing pangenomes, each tailored to suit specific research objectives and employing unique techniques. Minigraph generalizes minimap2, which only calls SVs (Li et al., 2020). Cactus uses a phylogenetic tree to guide the creation of multiple alignments (Armstrong et al., 2020), and the Cactus Pangenome Pipeline adapts Cactus to eliminate the need for a guide tree and adds base-level alignments to the minigraph graph, though it is still single reference-based (Hickey et al., 2023). PPanGGOLiN uses gene families as nodes and genomic neighborhoods as edges (Gautreau et al., 2020), and Pandora focuses on SNPs of pangenomes by constructing graphs based on individual multiple sequence alignments of coding sequences and intergenic regions (Colquhoun et al., 2021). Meanwhile, minimizer-space de Bruijn graphs offer a graph representation for highly accurate, long sequencing reads (Ekim et al., 2021). In contrast to these tools, the PanGenome Graph Builder (PGGB) stands out as the least unbiased method (Garrison et al., 2023). PGGB incorporates an "all-versus-all" alignment method, treating each input genome with equal importance. The graphs produced by PGGB provide a base-level representation of the pangenome, even within repetitive regions, and include variants of all scales, from SNPs to large SVs. This allows every included genome to serve as a reference for subsequent analysis (Garrison et al., 2023). PGGB has previously been used to build the draft human pangenome (Liao et al., 2023), and Guarracino and colleagues have used it to validate a longstanding hypothesis regarding the evolution of human acrocentric chromosomes (Guarracino et al., 2023). Therefore, we used PGGB for pangenome graph construction because of its comprehensive and unbiased capabilities.

The implementation of pangenome graphs in infectious disease research is crucial, offering significant advantages. The use of pangenome graphs not only allows for the identification of novel genetic variants and SVs that may be overlooked by traditional linear reference-based methods (Garrison et al., 2018), but also provides the potential to address some longstanding unresolved questions, such as the origin of antibiotic resistance (Forsberg et al., 2014), the evolution of pathogenicity (Zhou et al., 2020), and the impact of horizontal gene transfer and evolution of genome architecture (Soucy et al., 2015). *Neisseria (N.) meningitidis*, also known as the meningococcus pathogen, is the primary agent responsible for invasive meningococcal diseases such as meningitis and septicemia, causing isolated incidents, outbreaks, and epidemics worldwide (Halperin et al., 2012). The genome of this bacterium spans approximately 2.1–2.4 Mb and possesses a GC content ranging from 51%–52%. One striking characteristic of *Neisseria meningitidis* genomes is their high recombination rate, which largely fuels the extensive genetic diversity within this species (Schoen et al., 2009; Didelot and Maiden, 2010; Harrison et al., 2017). In this study, we utilized both real and simulated genomic data of *N. meningitidis*

to assess the pangenome pipeline, covering pangenome graph construction to variant calling. Our findings demonstrated that using pangenome graphs improves mapping rates and enhances variant calling. This heightened accuracy, encompassing all types of variants, has the potential to improve outbreak investigations, predict drug resistance, and facilitate vaccine design (Rasko et al., 2008; Naz et al., 2019). By employing the least unbiased pangenome graph construction tool PGGB and utilizing a graph reference for subsequent NGS data analysis, our pangenome graph pipeline offers a promising and practical approach for comparative genomics and comprehensive genetic variation analysis in infectious disease research. This paves the way for more accurate and in-depth investigations of pathogen diversity, evolution, and adaptation (Paten et al., 2017; Rakocevic et al., 2019).

Materials and methods

Background of *Neisseria meningitidis* NZMenB epidemic strain

In Aotearoa New Zealand (NZ), from 1991 to 2007, an extended serogroup B epidemic occurred due to a single strain known as NZMenB (designated B:4:P1.7-2,4), identified by the PorA variant (P1.7-2), which still accounts for around one-third of meningococcal disease cases in NZ (Dyet and Martin, 2006; Yang et al., 2021). Based on our unpublished WGS data, we have categorized NZMenB into three phylogenetic clades, namely, clade154, clade41 and clade42 based on the multilocus sequence types (MLST) of seven housekeeping genes for sequence type (ST), ST154, ST41 and ST42 respectively (Maiden et al., 1998). The epidemic was primarily driven by two monophyletic clades, namely, ST154 and ST42, which accounted for the majority of the disease cases. On the other hand, although fewer isolates were associated with ST41, it displayed greater diversity, with the presence of multiple distinct lineages.

Nanopore long-reads

To analyze the WGS dataset, the original reference genome NC_017518 (a ST42 isolate) was used. To obtain complete reference genomes for NMI01191 (a ST41 isolate) and NMI97348 (a ST154 isolate), we conducted Nanopore long-read sequencing. High molecular weight genomic DNA was extracted using the Gentra Puregene Yeast/Bact. Kit (QIAGEN) and purified with Agilent Magnetic Beads. We used 400 ng of high molecular weight genomic DNA to construct sequence libraries utilizing the SQK-RBK004 Rapid Barcoding kit (Oxford Nanopore Technologies). The libraries were sequenced on R9.4.1 MinION flow cells. We used Flye version 2.8.1 (Kolmogorov et al., 2019) for *de novo* assembly, and Illumina sequencing reads were employed to polish the assembly using Unicycler version 0.4.8 (Wick et al., 2017). Consequently, we were able to obtain complete NZMenB genomes (3STs) comprising NMI01191 for ST41, NMI97348 for ST154, and NC_017518 for ST42. The 3ST genomes were aligned using progressiveMauve (Darling et al., 2004).

Simulation of genomes for pangenome graph construction

Mauve alignments demonstrated large inversions among the 3ST genomes. To evaluate pangenome graph construction, we simulated three genomes from NC_017518 (ST42) by introducing either randomly generated SNPs or mutated according to the SNP differences of ST41 and ST154 relative to ST42. The simulation was followed by introducing 200 indels and two inversions using simuG (Yue and Liti, 2019). We named the three simulated genomes ST42Sim, ST41Sim, and ST154Sim. The three simulated genomes contained 200 indels and two inversions relative to ST42, with ST42Sim, ST41Sim and ST154Sim containing 5000, 2892 and 4283 SNPs respectively. We grouped the three simulated genomes with ST42, which we refer to as the 4Sim genomes, and used them for further analysis.

Downloading diverse *Neisseria meningitidis* genomes from NCBI

To expand our evaluation of pangenome graph construction to more diverse genomes, 130 *N. meningitidis* (NM) genomes were downloaded from NCBI (Supplementary Table S1). The 130NM genomes comprised 8, 20, 20, 62, 2, 13, and 5 of group A, B, C, W, X, Y, and ungrouped, respectively.

Pangenome graph construction with PGGB

We constructed pangenome graphs for the 4Sim genomes, 3STs of NZMenB, and 130NM genomes using the PanGenome Graph Builder (PGGB) (Garrison et al., 2023). PGGB is a reference-free method for graph construction by employing all-to-all alignments with wfmash, graph induction via seqwish, and progressive normalization using smoothxg and gfaffix, graph visualization and generating statistics using Optimized Dynamic Genome/Graph Implementation (ODGI) (Guarracino et al., 2022; Garrison et al., 2023). To construct the pangenome graphs, we initially aligned the start of ST41 and ST154 with ST42 for the 3STs, and all 130NM genomes were fixed to start with the *dna*A gene using circlator version 1.5.5 (Hunt et al., 2015).

There are three essential parameters for PGGB pangenome graph construction, -n, the number of genomes, -s, the segment length (defines the seed length for alignment used in wfmash), and -p, the minimum pairwise identity between seeds. Here, we explain how we optimized these parameters for our specific datasets. We adopted the mash triangle approach (Ondov et al., 2016) to estimate pairwise distances within each dataset. The maximum distance observed was 0.0038 for the 4Sim genomes, 0.0016 for the 3ST, and 0.0232 for the 130NM. Following the guidance provided by the PGGB developers, we slightly decreased the -p value in accordance with these pairwise distances for more inclusive all-to-all alignments with wfmash. When adjusting -s (1000, 2000, 5000, and 10000) and -p (96, 95 and 90) parameters for the 4Sim genomes, the resulting pangenome graphs were similar across the different parameter settings. Another parameter, -k, influences the graph structure significantly; it excludes matches shorter than a certain threshold

from the initial graph model, which we used the default -k 19. The PGGB developers suggest setting -k larger for larger genomes. Larger values for -k also allow us to ignore, when necessary, short homologies due to the intervention of transposable elements, which would increase the complexity of the graph. A lower -p value will result in more inclusive alignments, and a larger -s value can reduce graph complexity by focusing on longer homologies between the genomes being aligned. To finely adjust the PGGB tool for different datasets, these parameters (-s and -p) may require modification based on the specific properties of the genomes, such as their divergence and frequency of SVs. According to the divergence among genomes and known rearrangement in the dataset of 4Sim, 3ST, we set the parameters -s, -p, and -n to 1000, 96, and 4, respectively, for the 4Sim genomes, and to 2000, 95, and 3, respectively, for the 3ST genomes. For the 130NM genomes, we opted for a larger -s 10,000 value, both for scalability reasons and to keep graph complexity lower. As a result, we set the parameters -s, -p, and -n to 10000, 95, and 131, respectively. By employing these selected parameter values, we successfully generated the most concise pangenome graph for each dataset, guaranteeing the optimal alignment of a significant proportion of segments from each path within the graph (https://github.com/pangenome/pgge). Additionally, the "odgi stats -S" option was used to generate statistics for the seqwish and smoothxg graph and "multiqc -m" option was used to generate a MultiQC report of the graphs' statistics and visualizations. All runs were executed with 48 threads on a Dell R840 server with an Xeon Gold 6244 3.60 GHz CPU with 64 cores, and 3TB RAM at ESR. We also utilized gfaestus for the 2D visualization of the pangenome graphs of both the 4Sim and 130NM (https://github.com/chfi/gfaestus).

Vg deconstruct to call variants in the graphs

Variation graphs offer a compact representation of genetic variation across a population in the form of bidirected DNA sequence graphs, encompassing large-scale SVs like inversions and duplications (Paten et al., 2017). To identify both small and large variants from the pangenome graph, we employed the Variation Graph (VG) toolkit (Garrison et al., 2018) to deconstruct the variants into VCF files using the path NC_017518 (ST42). The VG toolkit, standing for Variation Graph, enables genomic analysis, such as alignment, assembly, and variant calling, directly on the graph structure, thereby facilitating the study of complex and highly variable genomic regions while maintaining the context of each variation's position in the genome. When employing the "vg deconstruct" feature, we set the parameters -a to process all snarls (genomic regions containing variant sites and corresponding alternative alleles), including nested ones, -e to consider traversals that correspond to paths in the graph, and -K to retain conflicted genotypes, thereby ensuring the inclusion of all variants present in the graph.

Given that the simulated genomes (ST42Sim, ST41Sim, and ST154Sim) were derived from ST42, the known variations for these simulated genomes relative to ST42 were served as the ground truth. By utilizing this ground truth information, we conducted a comparative analysis, evaluating the observed variations within the 4Sim genome graph. Initially, we filtered for variations larger than 100 bp, and then we utilized vcfallelicprimitives from vcflib v.1.0.0 (Garrison et al., 2022) to deconvolute complex variations that were less than 100 bp. We compared the variants identified in the graph with the established ground truth to evaluate their consistency. Variants were categorized as consistent if they were present in both the graph and the ground truth, as false negatives if they were present in the ground truth but not detected in the graph, and as false positives if they were found in the graph but not in the ground truth.

Simulated NGS dataset of *Neisseria meningitidis* for pangenome graph based variant calling

In addition to the comparative genomics analysis of the paths (genomes) based on the genome graphs, these graphs can also serve as a pangenome reference for NGS data analysis. To evaluate the genome graph-based pipeline for NGS data mapping and variant calling using the VG toolkit (Garrison et al., 2018), we simulated 100 × read depth 2 × 150 bp paired NGS data with an error rate of 0.5% using wgsim from samtools (Li et al., 2009).

We began with eight genomes, which included the 3ST genomes and the three simulated genomes, and two mutated genomes, ST41Mut and ST154Mut, based on the SNP difference of ST41 and ST154 relative to ST42. To generate a set of 40 genomes, we initially introduced 2000 SNPs for each of the eight genomes with five repeats, followed by two additional rounds of 2000 SNPs (40 genomes per round) using SimuG (Yue and Liti, 2019). Consequently, we obtained 128 genomes distributed among eight groups, including ST42, ST42Sim, ST41, ST41Mut, ST41Sim, ST154, ST154Mut, and ST154Sim.

Real NGS dataset of NZMenB for pangenome graph based variant calling

To test the graph-based analysis for a real NGS dataset, we mapped the NGS dataset of 149 NZMenB isolates to the 3ST pangenome graph (Supplementary Table S2). The 149 isolates included 49 from clade154, 48 from clade41 and 52 from clade42.

Map the NGS data to graph using the VG toolkit

To map the NGS data to genome graph using the VG toolkit, we initially converted graphs (4Sim and 3ST) into 256 bp chunks using the command "vg mod -X 256". We then employed "vg index" to generate the index for the graph. Subsequently, "vg map" was utilized to map the NGS data against the graph, resulting in the generation of gam files. We also used 'vg stats' to check the mapping statistics.

To compare the mapping rates for NGS dataset against linear references *versus* the graph, we also mapped the NGS data to each linear reference using Bowtie2 version 2.3.2 (Langmead and Salzberg, 2012). All sequenced and aligned reads were further processed using the Picard-tools version 2.10.10

(https://broadinstitute.github.io/picard/) to remove duplicated reads and were assessed with Qualimap version 2.2.1 (Garcia-Alcalde et al., 2012).

Variant calling for NGS data against genome graph

There are currently two popular approaches to call variants in pangenome graphs: genotyping known variants and novel variant calling. We utilized both methods to call variants for the 128 simulated NGS dataset against the 4Sim genome graph.

To genotype known variants in the graph, we employed "vg pack" to calculate the support reads for each gam file. We then utilized "vg call" to genotype the known variants for each sample based on the snarls file generated from the 4Sim genome graph.

To consider novel variants from the reads, we employed "vg augment" to augment each gam file. Subsequently, we indexed the augmented graph, calculated read support for all variants, and performed variant calling. High confidence variants were identified using the PASS information and genotype (GT = 1|1) from the VCF file. To evaluate the performance of variant calling in the context of simulated genomes, we compared the high confidence variants identified against the 4Sim graph with the simulated SNP records. As the actual variations of ST41 and ST154 relative to ST42 remain unknown, both sets of simulated NGS data were excluded from this analysis.

Distance matrices for cluster relationship

To analyze the cluster relationship among the 130NM genomes, we utilized "odgi similarity" from odgi version 0.8.3 (Guarracino et al., 2022) to extract a sparse similarity matrix for paths of the 130 MN graph. We then converted the paired Jaccard similarities from column six into a Jaccard distance matrix. These Jaccard similarities are measures that represent the proportion of shared elements between pairs of paths. We then employed hierarchical clustering to construct the phylogenetic relationship among the genomes based on the Jaccard distances. To assess the accuracy of the clustering relationship, we compared it to the one obtained by kmer-based SNP phylogenetic analysis.

For the kmer-based SNP analysis, we used ska, a reference-free, contig-based analysis, to extract the SNPs derived from default kmer length 31 that were present in 90% of the isolates (Harris, 2018). Phylogenetic analyses were constructed from the kmer-based SNP alignment using maximum likelihood under the best-fit model by Bayesian Information Criterion with iqtree version 2.0.6 (Minh et al., 2020). The robustness of the clades was estimated with 2000 ultra-fast bootstrap replicates.

Results

Overview of the pangenome graph workflow

A pangenome is defined as the comprehensive collection of whole-genome sequences from multiple individuals within a clade, a population or a species (Medini et al., 2005; Tettelin et al., 2005; Vernikos et al., 2015; Kavvas et al., 2018). This collective genomic dataset can be further divided into two distinct components: the core genome, which includes genes present in all individuals at the time of analysis, and the accessory genome, consisting of genes found only in a subset of individuals (Vernikos et al., 2015; Figure 1A). Pangenome graphs are pangenomes stored in graph models that can capture the entire genetic variation among genomes in a population or of a set of related organisms (Paten et al., 2017; Garrison et al., 2018; Eizenga et al., 2020; Garrison et al., 2023; Figure 1B).

In this study, we have developed a pangenome graph pipeline for microbial genomics, consisting of graph construction using PGGB (Garrison et al., 2023), graph manipulation through ODGI (Guarracino et al., 2022), and variant calling for NGS data using the VG toolkit (Garrison et al., 2018; Figure 1C). ODGI facilitates graph manipulation tasks such as visualization, and extraction of distances among paths in the graph, enabling phylogenetic analysis (Guarracino et al., 2022). Additionally, we utilized the VG toolkit for analyzing NGS data against the graph for read mapping and variant calling (Garrison et al., 2018).

The pangenome graph construction with PGGB was demonstrated to be effective across various datasets, though the resulting graphs varied significantly based on the complexity of the input genomes (Supplementary Table S3). The total run times for PGGB were 10.8 min, 8.3 min, and 4392 min, and the maximum memory usage was 1.87 GB, 2.01 GB, and 38.64 GB for the 4Sim, 3ST, and 130NM, respectively. In the case of the 130NM genomes, employing the PGGB tool with the "-x auto" option enabled for the giant component heuristic resulted in a total execution time of 2787 min and a peak memory usage of 21.92 GB. Notably, the generated graph remained identical to the one obtained without this option. In scenarios involving hundreds to thousands of genomes, it is recommended to utilize mapping sparsification (-x auto) to alleviate computational demands.

High consistency between variations in the 4Sim genome graph and ground truth

The final smoothed graph for the 4Sim genomes spanned 2,260,981 bp and consisted of 30,033 nodes and 40,273 edges. This is slightly larger than each of the input genomes: 2,248,966 bp for NC_017518 (ST42); 2,249,014 bp for ST41Sim, 2,248,965 bp for ST154Sim, and 2, 249,050 bp for ST42Sim. Mauve alignment (Figure 2A) supported our observations, as inversions were displayed as bubbles in the 2D visualization (Figure 2B) and as inverted sequences in the 1D visualization (Figure 2C). The VCF file indicated that inversions were identified as different genotypes across various genomes. It is important to note that some variations in the graph did not correspond to the ground truth due to alignment discrepancies in the indel regions (Figure 2D). Upon manual inspection of these sites, it was found that these variants represented the same variation but were aligned to either the left or the right of the indels in the graph compared to the ground truth. We detected four, three, and two false negative small variations for ST154Sim, ST41Sim, and ST42Sim, respectively, in comparison to ST42. Additionally, we identified seven false positive small variants in ST154Sim. Therefore, both sensitivity and specificity for variations in graph compared to ground truth are over 99.9%. Despite the relatively simple nature of the simulated genomes, the agreement between the variations in the graph and the ground truth implies that the pangenome graph

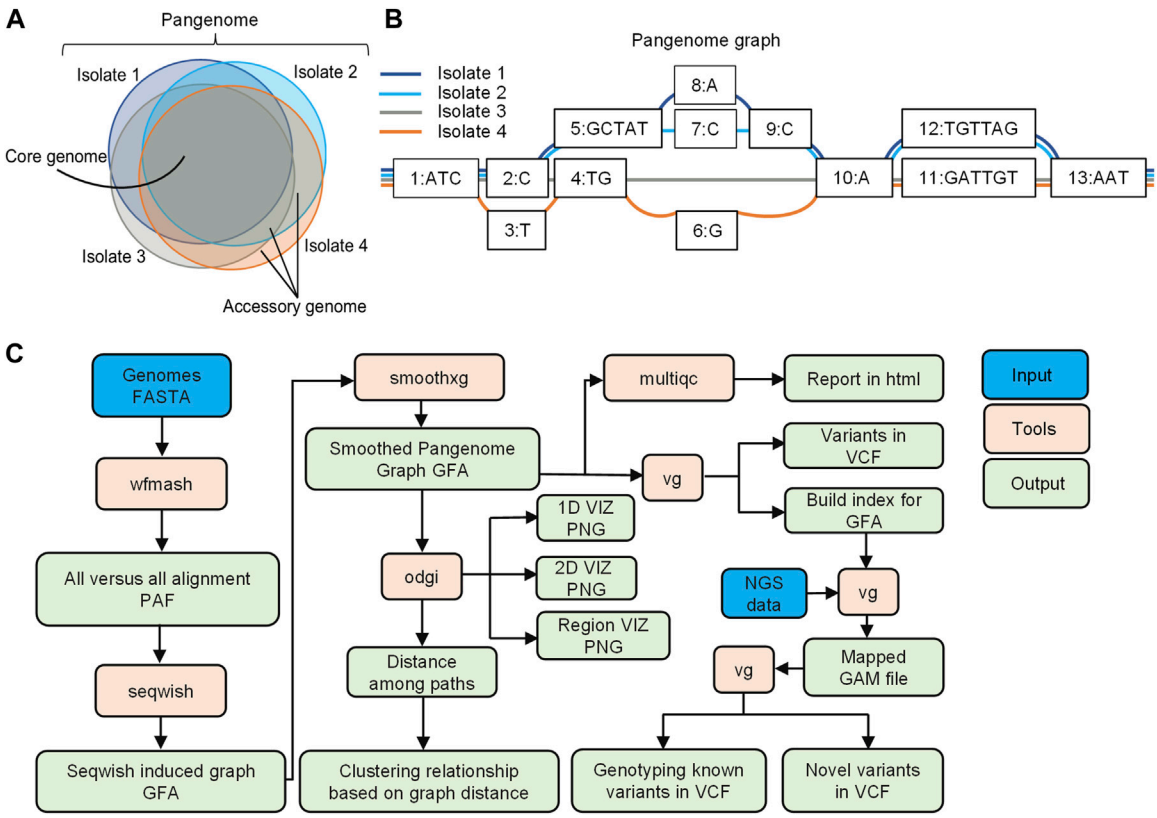


FIGURE 1
Overview of pangenome graph pipeline. **(A)** Bacterial pangenome, core genome, and accessory genomes. The pangenome represents the comprehensive collection of whole-genome sequences from multiple individuals within a clade. The core genomes comprise of a set of genes present in all individuals, while accessory genomes consist of genes found in a subset of individual genomes **(B)** Pangenome graphs representation. Pangenomes are stored in graph models, where nodes (numerically labeled) represent DNA segments of varying lengths. Edges connect nodes, and paths represent walks through the nodes of the graph, corresponding to the input genomes **(C)** Pangenome graph pipeline with PGGB. The pipeline includes graph construction using the PGGB tool, graph manipulation using ODGI, and variant calling for NGS data using the VG toolkit. The overview demonstrates an efficient and integrated approach to pangenome analysis.

generated by PGGB is able to accurately and reliably detect genetic variant across the input genomes (Supplementary Table S4).

100% mapping rates and comparable variant calling in graph-based analysis of simulated NGS data

Utilizing a pangenome graph reference for the analysis of NGS data can significantly enhance mapping rates. We conducted an evaluation by comparing the mapping rates of simulated NGS data based on the 4Sim graph to each of the linear references: ST42, ST42Sim, ST41Sim, and ST154Sim. All datasets were mapped to the graph, yielding a 100% mapping rate. Although the rates of NGS data aligned to each single linear reference were all over 99%, a bias was observed in the linear reference mapping rates (Figure 3A). Our findings indicate that the use of a pangenome graph reference can greatly improve mapping rates in NGS data analysis.

The pangenome graph integrates various genomic variants, making it possible to genotype variants in NGS datasets. Interestingly, the genotyped results demonstrated high consistency across the eight simulated NGS datasets (Figure 3B; Supplementary Table S6). The ST42Sim group exhibited the highest number of variants, consistent with the original simulation of 5000 SNPs and 200 indels. Moreover, the ST41Sim group displayed more identified variants compared to ST41 and ST42Mut, while the ST154Sim group revealed more variants compared to ST154 and ST154Mut.

Variant calling for NGS data against the graph using the VG toolkit differs slightly from single linear reference-based calling. In the absence of a defined path for variant calling, the process will call variants against the paths in alphabetical order (e.g., core genome part from A path, accessory genomes from B path, etc.). The variant call format file includes a PASS column to mark variants that pass all filters, and the genotype (GT) describes the identified genotype in each sample. Since we analysed haplotype bacterial genomes, variants with PASS but GT not equal to 1|1 were classified as errors, while those with PASS and GT = 1|1 were classified as high-confidence variants. For each simulated NGS group, high-confidence variants exhibited consistency. Interestingly, the ST41 and ST154 groups exhibited the lowest proportion of high-confidence variant calls, which may be attributed to these groups' greater genomic diversity and the absence of a reference from either group in the graph. Including one reference from each of these groups in the pangenome graph led to an improvement in the

10.3389/fgene.2023.1225248

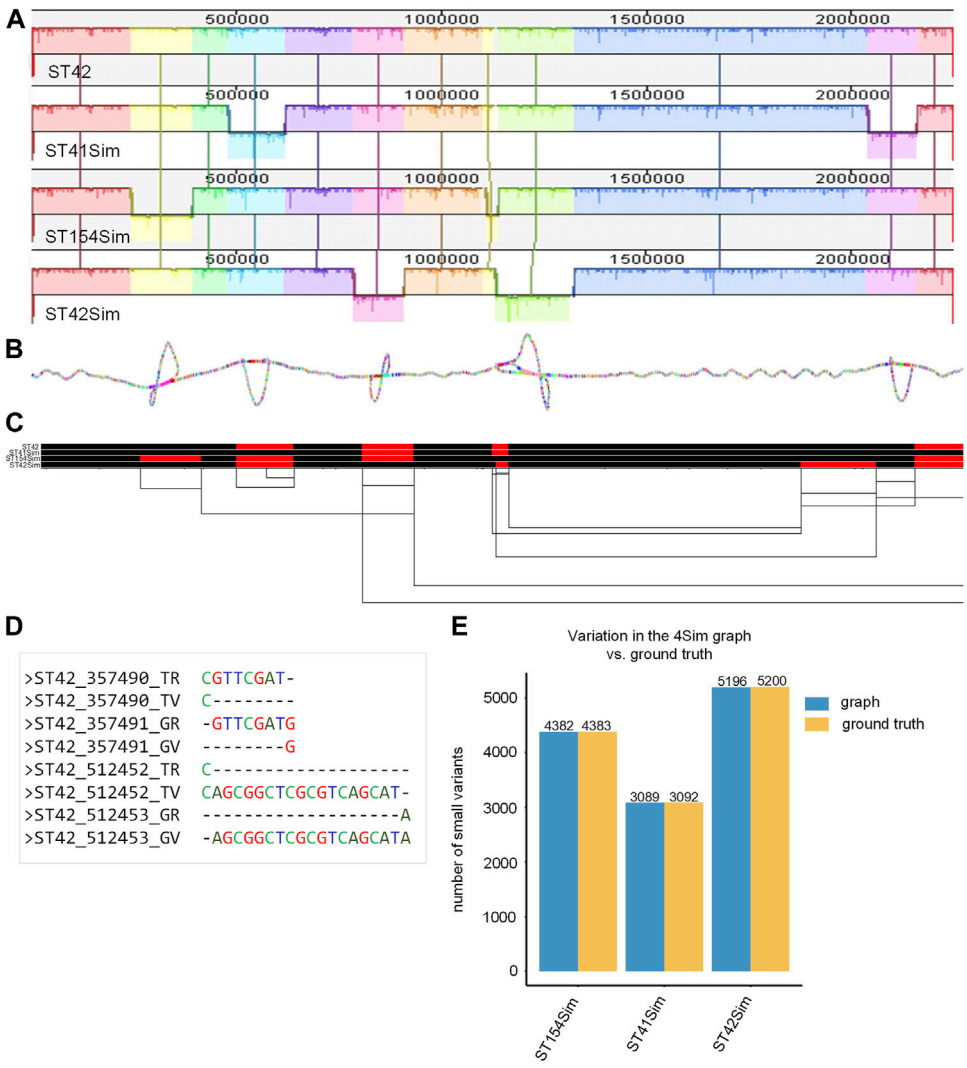


FIGURE 2
Pangenome graph of the 4Sim genomes. **(A)** Mauve alignment of the 4Sim Genomes. The Mauve alignment of the 4Sim genomes is depicted, with blocks under each line representing inverted regions **(B)** 2D visualization of the 4Sim pangenome graph. The pangenome graph of the 4Sim genomes is visualized in 2D using gfaestus. Bubbles in the graph indicate inversions. **(C)** 1D visualization of the 4Sim pangenome graph with path orientation, highlighting the inversions. The 4Sim pangenome graph is visualized in 1D using ODGI. Forward paths are represented in black, while reverse paths are in red. Regions displaying both black and red represent inversions **(D)** Inconsistency in indel region alignment: graph vs. ground truth. This panel provides two examples of inconsistent indel region alignment between the graph and the ground truth. For example, the deletion that appears at position 3547490 in ST42 according to the ground truth, is marked as being at position 3547491 in the graph. The labels are as follows: TR, true reference; TV, true variant; GR, reference in the graph; GV, variation in the graph. **(E)** Consistency of variations: graph vs. ground truth. A bar plot demonstrates the high consistency of variations between the graph and the ground truth, highlighting the accuracy of the pangenome graph representation.

proportion of high-confidence variant calls (Figure 3C; Supplementary Table S8). Furthermore, as NC_017518 (ST42) was the first path from the graph for variant calling, the ratio of high-confidence variants to the number of simulated variants was higher in ST42 (0.944–0.959) and ST42Sim (0.959–0.9706), but relatively lower in ST154Mut (0.8755–0.9000) and ST154Sim (0.8792–0.9049) (Figure 3D).

Enhanced mapping of NZMenB real NGS data to pangenome graph

The three sequence types (STs) represent the three major clades of NZMenB (Figure 4A). The final graph for 3STs spanned 2,304,073 bp, consisting of 23,323 nodes and 31,325 edges. This is marginally larger than each of the input genomes: 2,248,966 bp for NC_017518 (ST42); 2,217,832 bp for NMI01191 (ST41) and 2,233,582 bp for NMI97348 (ST154). The inverted regions are consistent in both the Mauve alignment (Figure 4B) and the 1D graph visualization (Figure 2C).

We mapped each group of genomes (ST154, ST41, and ST42) to the respective linear references - ST154, ST41, ST42, and the 3STs graph. Despite the higher diversity of the compared genomes, particularly within the ST41 group, the mapping rate was higher when mapped to the graph as opposed to a single linear reference (Figure 4D; Supplementary Table S9). For example, when examining the reads of ST154 and their mapping to the ST154, ST41, ST42, and 3ST genome graphs, we observed values ranging from 0.9721 to

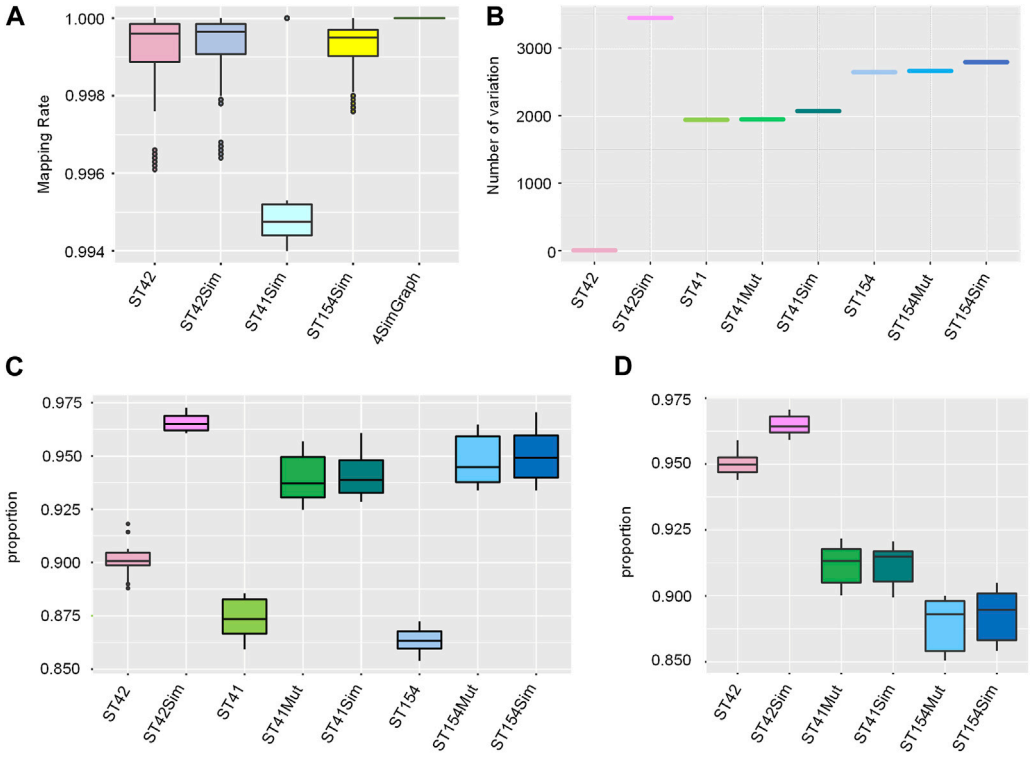


FIGURE 3
Mapping rates and comparable variant calling in graph-based analysis of simulated NGS Data. **(A)** Mapping rates: simulated NGS data to linear reference vs. 4Sim pangenome graph. This panel presents a comparison of mapping rates for simulated NGS data to each linear reference and the 4Sim pangenome graph. **(B)** Known variant genotyping in the 4Sim graph. A box plot displays the number of variations in genotyping based on known variants within the 4Sim graph. **(C)** Novel variants from graph-based calling. The box plot shows the proportion of high-confidence variants for each simulated group of data, illustrating the effectiveness of graph-based variant calling **(D)** Overlap of high-confidence variants with simulated variants. This box plot represents the proportion of high-confidence variants that overlap with simulated variants for each group, demonstrating the accuracy of graph-based variant calling in identifying true variations.

0.9973, 0.972 to 0.9967, 0.9738 to 0.998, and 0.9795 to 0.9999, respectively. The isolates of the ST154 group may be less diverse, as indicated by the smaller range of mapping rate variation, while the isolates of the ST41 group display greater diversity, as evidenced by the larger ranges of mapping rate variation (0.958–0.9956, 0.971 to 1, 0.9744 to 0.9988, and 0.9785 to 0.9998, respectively). The isolates belonging to the ST42 group displayed comparable mapping rates when mapped to both the ST42 and 3ST genome graphs. However, slightly lower mapping rates were observed when these isolates were mapped to ST154 (ranging from 0.9536 to 0.994) and ST41 (ranging from 0.9559 to 0.9959). In summary, these findings suggest potential reference bias when using a single linear reference and demonstrate that utilizing a graph as a reference can improve the mapping process.

The clustering relationships among paths in the genome graph effectively reveal phylogenetic connections

We evaluated the performance of the PGGB method on a diverse group of 130NM genomes, constructing a pangenome graph that proved more complex than the 4Sim and 3ST pangenomes. The 130NM pangenome graph spans 4,751,450 base pairs, over twice the size of a typical individual *N. meningitidis* genome and comprises 629,349 nodes and 894,725 edges.

The 1D visualization of the 130NM graph, which colours paths based on orientation, shows genome chunks as either forward (black) or reverse (red) (Figure 5A), illustrating the high recombination rate of *N. meningitidis* genomes. The 2D visualization using gafestus reveals large bubbles, potentially due to the substantial number of genomes aligned in reverse (Figure 5B). We classified the variations in the graph into (multiple) SNPs, indels and SVs. An example of a multiple nucleotide polymorphism (MNP) is when a sequence variation involves changes in multiple adjacent nucleotides. For example, a DNA sequence changes from "GGG" to "ACA". The 130NM pangenome graph contains 133, 745 (M) SNPs, 25,478 indels, and 1,446 SVs.

The all-vs-all alignment pangenome graph construction is unbiased, allowing distances among paths in the graph to effectively reveal genome relationships. Using the Jaccard similarity of the 130NM paths, we constructed a phylogenetic relationship among them. Clonal complexes are well-resolved by Jaccard similarity, with groups containing more than one genome clustering together (Figure 5C). This finding is largely consistent with phylogenetic relationships revealed by the kmer SNP-based analysis (Figure 5D). Most of the highly supported clades identified by the kmer SNP-based analysis are also clustered together on the

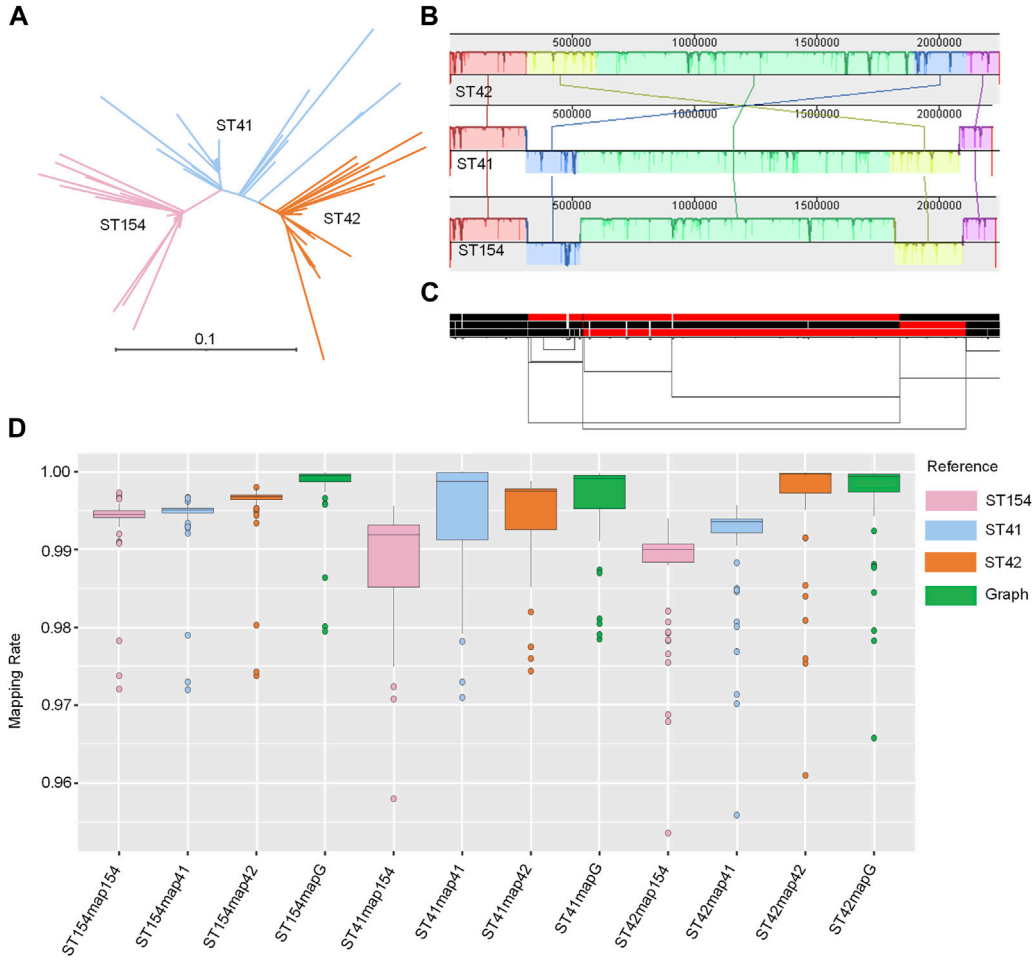


FIGURE 4
Pangenome graph of 3STs *N. meningitidis* genomes. **(A)** Phylogenetic analysis of NZMenB. The phylogenetic analysis of NZMenB reveals three major STs responsible for the epidemics: ST154, ST41, and ST42 **(B)** Mauve alignment of 3ST genomes. The Mauve alignment of the 3ST genomes is depicted, with blocks under each line representing inverted regions. **(C)** 1D visualization of the 3STs pangenome graph with path orientation. The 3STs pangenome graph is visualized in 1D using ODGI, displaying path orientation. Forward paths are represented in black, while reverse paths are in red. Regions displaying both black and red represent inversions **(D)** Mapping rates: real NZMenB NGS data to linear reference vs. 3STs to the pangenome graph. This panel presents a comparison of mapping rates for real NZMenB NGS data to each linear reference and the 3STs pangenome graph. Each group, ST154, ST41, and ST42, were mapped to their respective references and the graph.

Jaccard similarity tree, such as clonal complex ST8, ST23, ST175, ST420/6688, ST4, ST269, ST41/44, but the branches in the kmer SNP-based analysis are more diverse. There are two clonal complexes, ST-344 and ST-32, being clustered together on the Jaccard similarity tree but not on the kmer SNPs tree. Overall, the all-vs-all alignment pangenome graph is suitable for a relatively large number of genomes, capturing all types of variation and offering an unbiased method for genome comparison. The distance of pangenome graph paths reveals the genomic relationships well.

Discussion

Whole genome sequencing has revolutionized many aspects of infectious disease research, including the tracking and monitoring of pathogen spread and evolution (Didelot et al., 2012; Quick et al., 2016; Gardy and Loman, 2018; Geoghegan et al., 2020; Geoghegan et al., 2021; Yang et al., 2021), identification of drug susceptibility and resistance (Koser et al., 2014; Holt et al., 2015; Walker et al., 2015), and the advancement of vaccine development (Chen et al., 2021). However, the use of linear reference-based approaches for genomic analyses may lead to biases, particularly in studies focused on highly variable bacterial genomes (Darmon and Leach, 2014). To overcome the limitation of single linear reference genomes, pangenome graphs offer an efficient model for representing and analyzing multiple genomes and their variants within a graph structure encompassing all types of variations (Paten et al., 2017; Eizenga et al., 2020). In this study, we present a practical and unbiased bioinformatic pangenome graph pipeline (Figure 1C) that uses PGGB to construct pangenome graphs from assembled genomes for comparative genomics (Garrison et al., 2023), and employs the VG toolkit to align whole genome sequencing data and call variants against a graph reference (Garrison et al., 2018). We have demonstrated the efficacy of the pangenome pipeline across a diverse collection of *N. meningitidis* genomes, using both simulated and actual genomic datasets.

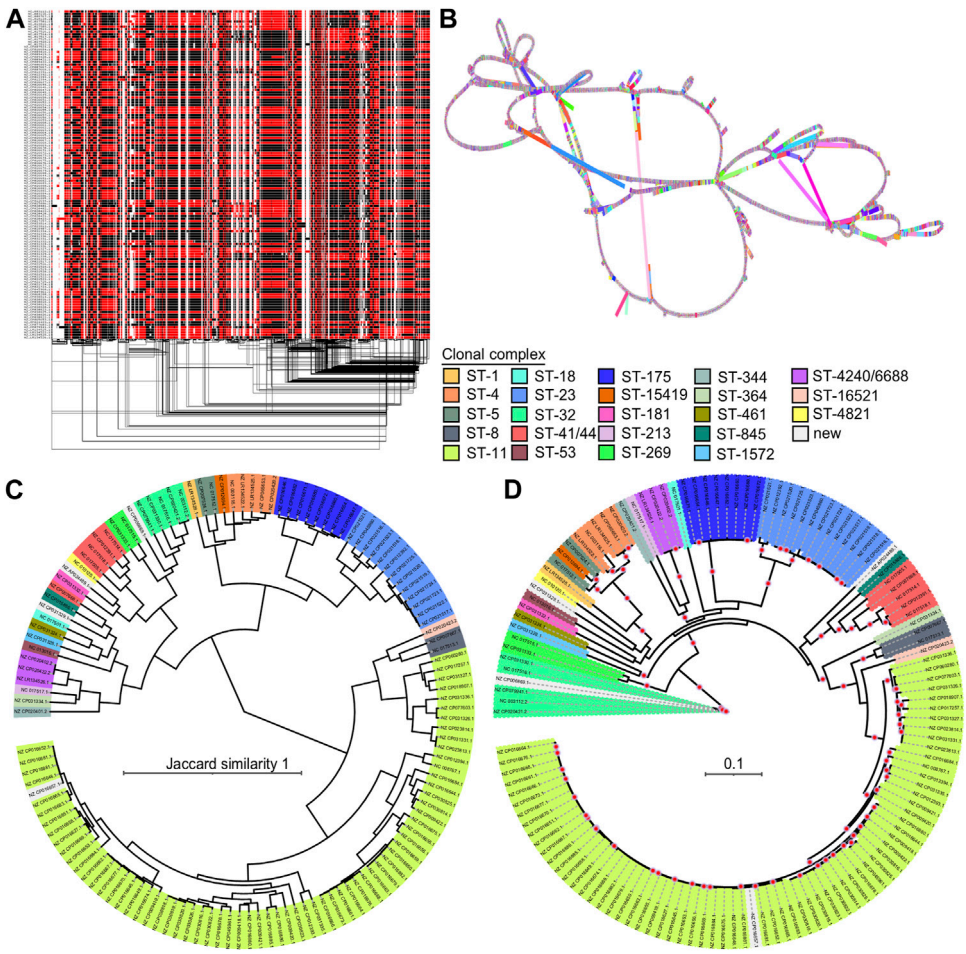


FIGURE 5
Pangenome graph of 130 *N. meningitidis* genomes and their phylogenetics relationships. **(A)** 1D visualization of the 130NM genomes with path orientation. The 130NM pangenome graph is visualized in 1D using ODGI, displaying path orientation. Forward paths are represented in black, while reverse paths are in red. Regions displaying both black and red represent inversions **(B)** 2D visualization of 130NM pangenome graph. The pangenome graph of the 130NM genomes is visualized in 2D using gfaestus. **(C)** Phylogenetic analysis of 130 NM genomes based on Jaccard distance of paths. The clustering relationship of 130 NM genomes is conducted based on Jaccard distance of paths in the 130NM pangenome graph. Isolate names and clades are coloured according to their clonal complex designation, with "New" indicating isolates where the clonal complex is not yet assigned **(D)** Phylogenetic analysis of 130NM genomes based on kmer SNPs. A maximum-likelihood phylogeny is constructed using iqtree v.2.0.6 under the best-fit model determined by the Bayesian Information Criterion. Branches with greater than 95% bootstrap consensus (from 2000 ultra-fast bootstrap replicates) are highlighted with a red dot. Isolate names and clades are coloured according to their clonal complex designation.

Numerous methods exist for constructing pangenomes, each with specific strengths and strategies (Armstrong et al., 2020; Gautreau et al., 2020; Li et al., 2020; Colquhoun et al., 2021; Ekim et al., 2021; Hickey et al., 2023); however, the PanGenome Graph Builder (PGGB) distinguishes itself by providing a comprehensive, unbiased approach that includes all types of genomic variations and treats each input genome equally (Guarracino et al., 2023). Using PGGB, we have successfully constructed pangenome graph for diverse datasets of *Neisseria meningitis* (the 4Sim, the 3STs and 130NM datasets). The resulting graphs varied considerably based on input genome complexity (Supplementary Table S3). The resulting pangenome graph proved to be a powerful tool for visualizing and analyzing the complex genomic relationships among these highly recombinant *Neisseria* genomes (Figures 2, 4, 5). By capturing all types of genomic variation and enabling unbiased genome comparisons, this approach offers significant advantages for comparative genomics studies. The accurate representation of inversions, SNPs, and indels in the graph for the 4Sim genomes (Figure 2; Supplementary Table S4) serves as strong evidence for the effectiveness of PGGB. Moreover, the flexibility offered by PGGB to adjust parameters according to the user's dataset is noteworthy. When using PGGB for pangenome graph construction, one can specifically adjust essential parameters such as -n, -s, and -p. These adjustments provide tuning opportunities to generate optimized graphs for different input datasets (Guarracino et al., 2023). In addition, enabling the -x auto option, the heuristic based on a model of random graphs to set a sparsification threshold for initial mappings of the 130NM dataset can significantly reduce computational time and maximum memory usage, but without compromising the accuracy of the final pangenome graph results (Supplementary Table S3).

In addition to representing various types of genomic variation in the pangenome graph generated by PGGB, we can also utilize distance metrics, such as the Jaccard distance of paths in the graph, to examine genomic relationships. Strains of *N. meningitidis* were classified into distinct clonal complexes based on similarity of STs by MLST (Maiden et al., 1998), reflecting their close evolutionary relationships. However, the high recombination rate of meningococcal genomes complicates the interpretation of phylogenetic relationships among strains and clonal complexes, and there is a need for novel genomic approaches to better understand their evolution (Harrison et al., 2017). For the diverse 130NM genomes, most highly supported clades identified by the kmer SNP-based analysis were also clustered together on the Jaccard distance tree (Figures 5C, D). This consistency underscores the utility of the pangenome graph approach for uncovering the underlying genomic relationships among *N. meningitidis* strains. Interestingly, we observed that the branches in the kmer SNP-based analysis are more diverse, suggesting that combining different methods of analysis may provide a more comprehensive understanding of the phylogenetic relationships among clonal complexes.

To circumvent reference bias, utilizing a pangenome as a reference is a significant direction for future genomics studies. In addition to pangenome graph construction using the PGGB method, our pipeline also employs the VG toolkit for the analysis of NGS data, which includes mapping and variant calling. Both simulated NGS and real data demonstrate improved mapping rates when using graph-based references compared to linear references, indicating that the adoption of a pangenome graph reference can substantially enhance mapping rates in NGS data analysis (Figure 3A; Figure 4D). The pangenome graph effectively integrates various genomic variants, enabling the genotyping of variants in NGS datasets that cannot be achieved using a single linear reference (Figure 3B). Furthermore, the novel variant calling approach based on the graph provides increased flexibility, allowing for either pangenome-based or reference-based variant calling. This feature significantly reduces reference bias and improves data analysis efficiency. Our results also reveal that the proportion of novel variant calls is remarkably high (Figure 3C), and a large number of simulated variations are identified (Figure 3D), indicating the reliability of graph-based NGS data analysis.

The incorporation of unbiased pangenome graphs into infectious disease research represents a remarkable advancement, yielding numerous benefits. Our pipeline employs PGGB for pangenome construction, which treats all input genomes in tandem, giving equal importance to every base. This comprehensive approach allows us to discern all genetic variation particularly structural variation and copy number variation that were likely overlooked by previous methodologies based on the use of a single reference genome. This enhanced detection capability proves crucial for the identification of virulence and antimicrobial resistance genes (Ekim et al., 2021). Simultaneously including all variations enhances our understanding of the genomes' evolutionary history, helping elucidate transmission patterns and establish connections between cases. This could prove invaluable in infectious disease research, where identifying the source or potential origins of new outbreaks is a priority, rapid genotyping against a graph could offer essential clues. Moreover, with multiple genomes integrated into the graph, each genome or the entire pangenome can serve as a reference for novel variant calling. This feature becomes especially valuable in public health surveillance, eliminating the need to try different references. This unbiased pangenome graph approach holds the potential to address longstanding challenges in infectious disease research, such as the origin of antibiotic resistance, a concern with significant public health implications. Pangenome graphs can be used to track and understand the genetic determinants contributing to resistance, providing insights that could guide the selection of therapeutic modalities or preventive strategies. They allow researchers to visualize and interpret the complex genetic interactions and variations that propel the evolution of pathogenicity, thereby fostering a deeper understanding of pathogen behavior (Zhou et al., 2020). Additionally, they can elucidate processes such as horizontal gene transfer and evolution of genome architecture, both crucial for bacterial adaptability and survival (Soucy et al., 2015). Pangenome graphs are particularly beneficial for viral genomics studies, as viral genomes are smaller. The unbiased analysis of these genomes could provide evidence about their origin and spread, guiding the design of better vaccinations, and enhancing our ability to prevent, monitor, and treat infections.

Although the concept of pangenomes initially emerged from microbial research (Medini et al., 2005; Tettelin et al., 2005; Vernikos et al., 2015; Kavvas et al., 2018), pangenome graphs have since been applied to various species, such as humans (Guarracino et al., 2023; Liao et al., 2023), and cattle (Talenti et al., 2022). Integrating all genomic variants facilitates a comprehensive and unbiased view of the genetic landscape, as demonstrated by the draft human pangenome that not only captures known variants, haplotypes, and new alleles at complex loci but also adds 119 million base pairs of polymorphic sequences and 1,115 gene duplications compared to the existing GRCh38 reference (Liao et al., 2023). The research conducted by Guarracino et al. (2023) using PGGB confirmed a long-held hypothesis regarding the evolution of human acrocentric chromosomes—that these chromosomes contain pseudo-homologous regions where heterologous pairs recombine as if they were homologs. Pangenome graphs hold potential in the broader field of genomics, including human genetics and personalized medicine, where they can help uncover subtle genetic variations associated with disease susceptibility or treatment response. These methods are also expected to find applications in metagenomics, transcriptomics, and epigenomics, aiding in a more comprehensive understanding of genomic diversity.

In conclusion, the current pangenome pipeline has several advantages over other pipelines, offering a more comprehensive and accurate approach for comparative genomics and comprehensive genetic variation analysis for infectious disease. Pangenome graphs provide a promising and practical approach for advancing our understanding of pathogen diversity, evolution, and adaptation.

Data availability statement

The 4Sim, 3ST, 130NM datasets used for pangenome graph construction, 128 simulates genomes, and the scripts used in this

study are available at https://github.com/ZoeYang2020/Pangenome-Graphs-in-Infectious-Disease. 149 isolates of NZmenB NGS data are available for download from the National Center for Biotechnology Information Sequence Read Archive (Bioproject accession no. PRJNA592848 and PRJEB28859). All the related materials are already publicly available.

Author contributions

The project was designed by JdL and MB. DNA extraction, library construction, and sequencing experiments were carried out by ZY. Guidance on utilizing PGGB, ODGI and the VG toolkits for the pangenome graph pipeline was provided by AG and EG. ZY and NI conducted the data analysis. ZY authored the manuscript, which was critically revised and edited by JW, JdL, MB, PB, TM, EG, and PP. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by Genomics Aotearoa (Genome Graphs project to JdL and MB) via funding from the Ministry of Business Innovation and Employment (MBIE), ESR Strategic Science Investment Fund from MBIE (Pangenome Graph for Comprehensive Surveillance of Plasmids Conferring Antimicrobial Resistance in Aotearoa New Zealand to ZY). The N. meningitidis nanopore and NGS sequencing was made possible through funding from Health Research Council grant no. 17-364, awarded to Philip Carter and XR.

Acknowledgments

We extend our sincere appreciation to Dr. Xiaoyun Ren and Heather Davis for her invaluable assistance in culturing isolates for DNA sequencing. We also thank Russell Smith for his expertise in installing software and maintaining the HPC server at ESR. In addition, we express our gratitude to Philip Carter, Anjan Purkayastha and Tania Grieve for their insightful comments and feedback.

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.2023.1225248/full#supplementary-material

References

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 587 (7833), 246–251. doi:10.1038/s41586-020-2871-y

Chen, J., Wang, J., Zhang, J., and Ly, H. (2021). Advances in development and application of influenza vaccines. *Front. Immunol.* 12, 711997. doi:10.3389/fimmu.2021.711997

Colquhoun, R. M., Hall, M. B., Lima, L., Roberts, L. W., Malone, K. M., Hunt, M., et al. (2021). Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biol.* 22 (1), 267. doi:10.1186/s13059-021-02473-1

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14 (7), 1394–1403. doi:10.1101/gr.2289704

Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78 (1), 1–39. doi:10.1128/MMBR.00035-13

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., and Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* 13 (9), 601–612. doi:10.1038/nrg3226

Didelot, X., and Maiden, M. C. J. (2010). Impact of recombination on bacterial evolution. *Trends Microbiol.* 18 (7), 315–322. doi:10.1016/j.tim.2010.04.002

Domingo, E., Garcia-Crespo, C., and Perales, C. (2021). Historical perspective on the discovery of the quasispecies concept. *Annu. Rev. Virol.* 8 (1), 51–72. doi:10.1146/annurev-virology-091919-105900

Dyet, K. H., and Martin, D. R. (2006). Clonal analysis of the serogroup B meningococci causing New Zealand's epidemic. *Epidemiol. Infect.* 134 (2), 377–383. doi:10.1017/S0950268805004954

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., et al. (2020). Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* 21, 139–162. doi:10.1146/annurev-genom-120219-080406

Ekim, B., Berger, B., and Chikhi, R. (2021). Minimizer-space de Bruijn graphs: whole-genome assembly of long reads in minutes on a personal computer. *Cell Syst.* 12 (10), 958–968.e6. doi:10.1016/j.cels.2021.08.009

Forsberg, K. J., Patel, S., Gibson, M. K., Lauber, C. L., Knight, R., Fierer, N., et al. (2014). Bacterial phylogeny structures soil resistomes across habitats. *Nature* 509 (7502), 612–616. doi:10.1038/nature13377

Garcia-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Gotz, S., Tarazona, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28 (20), 2678–2679. doi:10.1093/bioinformatics/bts503

Gardy, J. L., and Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* 19 (1), 9–20. doi:10.1038/nrg.2017.88

Garrison, E., Guarracino, A., Simon, H., Flavia, V., Bao, Z., Tattini, L., et al. (2023). *Building pangenome graphs*. https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1.

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., and Prins, P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *Plos Comput. Biol.* 18 (5), e1009123. doi:10.1371/journal.pcbi.1009123

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Genomics*. https://arxiv.org/abs/1207.3907.

Garrison, E., Siren, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36 (9), 875–879. doi:10.1038/nbt.4227

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., et al. (2020). PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *Plos Comput. Biol.* 16 (3), e1007732. doi:10.1371/journal.pcbi.1007732

Geoghegan, J. L., Douglas, J., Ren, X. Y., Storey, M., Hadfield, J., Silander, O. K., et al. (2021). Use of genomics to track coronavirus disease outbreaks, New Zealand. *Emerg. Infect. Dis.* 27 (5), 1317–1322. doi:10.3201/eid2705.204579

Geoghegan, J. L., Ren, X. Y., Storey, M., Hadfield, J., Jelley, L., Jefferies, S., et al. (2020). Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat. Commun.* 11 (1), 6351. doi:10.1038/s41467-020-20235-8

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17 (6), 333–351. doi:10.1038/nrg.2016.49

Guarracino, A., Buonaiuto, S., Lima, L., Potapova, T., Rhie, A., Koren, S., et al. (2023). Recombination between heterologous human acrocentric chromosomes. *Nature* 617 (7960), 335–343. doi:10.1038/s41586-023-05976-y

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (2022). Odgi: understanding pangenome graphs. *Bioinformatics* 38 (13), 3319–3326. doi:10.1093/bioinformatics/btac308

Halperin, S. A., Bettinger, J. A., Greenwood, B., Harrison, L. H., Jelfs, J., Ladhani, S. N., et al. (2012). The changing and dynamic epidemiology of meningococcal disease. *Vaccine* 30, B26–B36. doi:10.1016/j.vaccine.2011.12.032

Harris, S. R. (2018). *Ska: Split kmer analysis toolkit for bacterial genomic epidemiology*. Available at: https://www.biorxiv.org/content/10.1101/453142v1.

Harrison, O. B., Schoen, C., Retchless, A. C., Wang, X., Jolley, K. A., Bray, J. E., et al. (2017). Neisseria genomics: current status and future perspectives. *Pathog. Dis.* 75 (6), ftx060. doi:10.1093/femspd/ftx060

Hickey, G., Monlong, J., Ebler, J., Novak, A., Eizenga, J. M., Gao, Y., et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* doi:10.1038/s41587-023-01793-w

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. U. S. A.* 112 (27), E3574–E3581. doi:10.1073/pnas.1501049112

Hunt, M., De Silva, N., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* 16, 294. doi:10.1186/s13059-015-0849-0

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., et al. (2018). Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* 9, 4306. doi:10.1038/s41467-018-06634-y

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37 (5), 540–546. doi:10.1038/s41587-019-0072-8

Koser, C. U., Ellington, M. J., and Peacock, S. J. (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.* 30 (9), 401–407. doi:10.1016/j.tig.2014.07.003

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi:10.1038/nmeth.1923

Li, H., Feng, X. W., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 21, 265. doi:10.1186/s13059-020-02168-z

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., et al. (2023). A draft human pangenome reference. *Nature* 617, 312–324. doi:10.1038/s41586-023-05896-x

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395 (10224), 565–574. doi:10.1016/S0140-6736(20)30251-8

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95 (6), 3140–3145. doi:10.1073/pnas.95.6.3140

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15 (6), 589–594. doi:10.1016/j.gde.2005.09.006

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37 (5), 1530–1534. doi:10.1093/molbev/msaa015

Mohd, H. A., Al-Tawfiq, J. A., and Memish, Z. A. (2016). Middle East Respiratory Syndrome Coronavirus (MERS-CoV) origin and animal reservoir. *Virol. J.* 13, 87. doi:10.1186/s12985-016-0544-0

Naz, K., Naz, A., Ashraf, S. T., Rizwan, M., Ahmad, J., Baumbach, J., et al. (2019). PanRV: pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *BMC Bioinform* 20 (1), 123. doi:10.1186/s12859-019-2713-9

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376 (6588), 44–53. doi:10.1126/science.abj6987

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 132. doi:10.1186/s13059-016-0997-x

Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res.* 27 (5), 665–676. doi:10.1101/gr.214155.116

Plowright, R. K., Parrish, C. R., McCallum, H., Hudson, P. J., Ko, A. I., Graham, A. L., et al. (2017). Pathways to zoonotic spillover. *Nat. Rev. Microbiol.* 15 (8), 502–510. doi:10.1038/nrmicro.2017.45

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530 (7589), 228–232. doi:10.1038/nature16996

Rakocevic, G., Semenyuk, V., Lee, W. P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* 51, 354–362. doi:10.1038/s41588-018-0316-4

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* 190 (20), 6881–6893. doi:10.1128/JB.00619-08

Sanjuán, R., and Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cell Mol. Life Sci.* 73 (23), 4433–4448. doi:10.1007/s00018-016-2299-6

Schoen, C., Tettelin, H., Parkhill, J., and Frosch, M. (2009). Genome flexibility in *Neisseria meningitidis*. *Vaccine* 27 (2), B103–B111. doi:10.1016/j.vaccine.2009.04.064

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16 (8), 472–482. doi:10.1038/nrg3962

Talenti, A., Powell, J., Hemmink, J. D., Cook, E. A. J., Wragg, D., Jayaraman, S., et al. (2022). A cattle graph genome incorporating global breed diversity. *Nat. Commun.* 13 (1), 910. doi:10.1038/s41467-022-28605-0

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome. *Proc. Natl. Acad. Sci. U. S. A.* 102 (39), 13950–13955. doi:10.1073/pnas.0506758102

The Computational Pan-Genomics Consortium (2018). Computational pan-genomics: status, promises and challenges. *Brief. Bioinform* 19 (1), 118–135. doi:10.1093/bib/bbw089

Vernikos, G., Medini, D., Riley, D. R., and Ttelin, H. (2015). Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154. doi:10.1016/j.mib.2014.11.016

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., et al. (2015). Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* 15 (10), 1193–1202. doi:10.1016/S1473-3099(15)00062-6

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 (6), e1005595. doi:10.1371/journal.pcbi.1005595

Xu, R. H., He, J. F., Evans, M. R., Peng, G. W., Field, H. E., Yu, D. W., et al. (2004). Epidemiologic clues to SARS origin in China. *Emerg. Infect. Dis.* 10 (6), 1030–1037. doi:10.3201/eid1006.030852

Yang, Z. Y., Ren, X. Y., Davies, H., Wood, T., Lopez, L., Sherwood, J., et al. (2021). Genomic surveillance of a globally circulating distinct group W clonal complex 11 meningococcal variant, New Zealand, 2013–2018. *Emerg. Infect. Dis.* 27 (4), 1087–1097. New Zealand, 2013-2018. doi:10.3201/eid2704.191716

Yue, J. X., and Liti, G. N. (2019). simuG: a general-purpose genome simulator. *Bioinformatics* 35 (21), 4442–4444. doi:10.1093/bioinformatics/btz424

Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y., and Achtman, M. (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia* pestis phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* 30 (1), 138–152. doi:10.1101/gr.251678.119

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 24 August 2023
DOI 10.3389/fmicb.2023.1216008

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Laura Berneking,
The Rockefeller University, United States
Yuan Min,
Chinese Center For Disease Control and Prevention, China

*CORRESPONDENCE
José E. Villacís
✉ jevillacis@gmail.com
Monica C. Gestal
✉ monica.cartellegestal@lsuhs.edu;
✉ mcarges@gmail.com

†These authors have contributed equally to this work

RECEIVED 03 May 2023
ACCEPTED 01 August 2023
PUBLISHED 24 August 2023

CITATION
Villacís JE, Castelán-Sánchez HG, Rojas-Vargas J, Rodríguez-Cruz UE, Albán V, Reyes JA, Meza-Rodríguez PM, Dávila-Ramos S, Villavicencio F, Galarza M and Gestal MC (2023) Emergence of *Raoultella ornithinolytica* in human infections from different hospitals in Ecuador with *OXA*-48-producing resistance. *Front. Microbiol.* 14:1216008. doi: 10.3389/fmicb.2023.1216008

COPYRIGHT
© 2023 Villacís, Castelán-Sánchez, Rojas-Vargas, Rodríguez-Cruz, Albán, Reyes, Meza-Rodríguez, Dávila-Ramos, Villavicencio, Galarza and Gestal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Emergence of *Raoultella ornithinolytica* in human infections from different hospitals in Ecuador with *OXA*-48-producing resistance

José E. Villacís [1,2*†], Hugo G. Castelán-Sánchez [3,4†], Jorge Rojas-Vargas [5†], Ulises E. Rodríguez-Cruz [6†], Viviana Albán [2,7†], Jorge A. Reyes [8], Pablo M. Meza-Rodríguez [4], Sonia Dávila-Ramos [4], Fernando Villavicencio [2], Margarita Galarza [9] and Monica C. Gestal [10*]

[1]Centro de Investigación para la Salud en América Latina (CISeAL), Pontificia Universidad Católica del Ecuador, Quito, Ecuador, [2]Centro de Referencia Nacional de Resistencia a los Antimicrobianos, Instituto Nacional de Investigación en Salud Pública, "Leopoldo Izquieta Pérez," Quito, Ecuador, [3]Programa Investigadoras e Investigadores por México, Grupo de Genómica y Dinámica Evolutiva de Microorganismos Emergentes, Consejo Nacional de Ciencia y Tecnología, México City, Mexico, [4]Centro de Investigación en Dinámica Celular, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico, [5]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [6]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México City, Mexico, [7]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, United States, [8]Facultad de Ciencias Químicas, Universidad Central del Ecuador, Quito, Ecuador, [9]Hospital de Especialidades Eugenio Espejo, Quito, Ecuador, [10]Department of Microbiology and Immunology, Louisiana State University (LSU), Health Science Center at Shreveport, Shreveport, LA, United States

Purpose: The purpose of this study was to highlight the clinical and molecular features of 13 *Raoultella ornithinolytica* strains isolated from clinical environments in Ecuador, and to perform comparative genomics with previously published genomes of *Raoultella* spp. As *Raoultella* is primarily found in environmental, clinical settings, we focused our work on identifying mechanisms of resistance that can provide this bacterium an advantage to establish and persist in hospital environments.

Methods: We analyzed 13 strains of *Raoultella ornithinolytica* isolated from patients with healthcare associated infections (HAI) in three hospitals in Quito and one in Santo Domingo de Los Tsáchilas, Ecuador, between November 2017 and April 2018. These isolates were subjected to phenotypic antimicrobial susceptibility testing, end-point polymerase chain reaction (PCR) to detect the presence of carbapenemases and whole-genome sequencing.

Results: Polymerase chain reaction revealed that seven isolates were positive isolates for bla_{OXA-48} and one for bla_{KPC-2} gene. Of the seven strains that presented the bla_{OXA-48} gene, six harbored it on an IncFII plasmid, one was inserted into the bacterial chromosome. The bla_{KPC} gene was detected in an IncM2/IncR plasmid. From the bioinformatics analysis, nine genomes had the gene bla_{OXA-48}, originating from Ecuador. Moreover, all *R. ornithinolytica*

strains contained the ORN-1 gene, which confers resistance for β-lactams, such as penicillins and cephalosporins. Comparative genome analysis of the strains showed that the pangenome of *R. ornithinolytica* is considered an open pangenome, with 27.77% of core genes, which could be explained by the fact that the antibiotic resistance genes in the ancestral reconstruction are relatively new, suggesting that this genome is constantly incorporating new genes.

Conclusion: These results reveal the genome plasticity of *R. ornithinolytica*, particularly in acquiring antibiotic-resistance genes. The genomic surveillance and infectious control of these uncommon species are important since they may contribute to the burden of antimicrobial resistance and human health.

KEYWORDS

***Raoultella ornithinolytica*, antimicrobial resistance (AMR), pangenome analyses, Ecuador (country), whole genome sequencing (WGS)**

1. Introduction

Raoultella species are gram-negative encapsulated bacilli belonging to Enterobacteriaceae (Drancourt et al., 2001). Until 2001, *Raoultella* species were considered part of the genus *Klebsiella*; however, with the current advances in molecular analysis based on *rpoB* sequencing, *Raoultella* was classified as a distinct and unique genus (Boye and Hansen, 2003; Martínez et al., 2004). Species such as *R. terrigena*, *R. planticola*, *R. electrica*, *R. trevisani*, and *R. ornithinolytica* belong to this genus. *R. ornithinolytica* is the most important because it has been associated with symptomatic cases of bacteremia (Kanki et al., 2002; Mau and Ross, 2010; Haruki et al., 2014; Tafoukt et al., 2017; Avellaneda et al., 2020; Wyres et al., 2020), urinary tract infections (García-Lozano et al., 2013; Haruki et al., 2014), joint infections (Beye et al., 2018), and biliary tract infections (Cleveland et al., 2014).

Raoultella ornithinolytica, is considered an unusual microorganism in health settings (Reyes et al., 2020). These bacteria are environmental microorganisms found in water, soil, and plants. For many years, this species have not been considered harmful to humans (Kanki et al., 2002). However, it has been found that some *R. ornithinolytica* strains may harbor antibiotic-resistance genes, such as $bla_{\text{NDM}-1}$, $bla_{\text{OXA}-48}$, and $bla_{\text{OXA}-232}$ in the environment (Iovleva et al., 2019), constituting a possible route of transmission and spread of antimicrobial resistance genes (ARGs) through mobile elements (horizontal gene transfer) (Tafoukt et al., 2017; Weng et al., 2020; Zou et al., 2020). A recent report from Croatia, described a case of septicemia in a 64-year-old male patient, caused by *R. ornithinolytica* and *Klebsiella pneumoniae*, both associated with antibiotic resistance and presence of $bla_{\text{OXA}-48}$ gene, which contributed to severity of infection and course of antibiotic treatment.

The correct identification of *R. ornithinolytica* is one of the main challenges in clinical settings. *R. planticola* and *R. ornithinolytica* share from of 98.3 to 99.5% of their genome content, which leads to form a tight phylogenetic cluster (Bravo, 2006; Walckenaer et al., 2008). Since most clinical laboratories relie on routine automated systems, high rates of misidentification have been reported, (Park et al., 2011; Sękowska et al., 2018). These systems are not sensitive enough to discriminate one species from the other through the ornithine decarboxylase (ODC) assay (Alves et al., 2006). In the absence of a robust biochemical assay and specific genetic markers allowing to detect of differences between these two species, the application of proteomic and genomic tools with whole-genome sequencing (WGS) have aided to accurately and rapidly discriminate at the species level. WGS allows us to have a greater number of genetic elements to differentiate them, but also to assess or infer functionality in terms of antimicrobial resistance (Galata et al., 2018).

In this study, we performed a comparative genomic analysis of 13 *R. ornithinolytica* strains isolated from different Ecuadorian hospitals identified by Whole Genome Sequencing (WGS). This comparative analysis revealed that the 13 bacterial strains correspond to different upsurge.

2. Materials and methods

2.1. Bacterial strains

We received 13 *Raoultella* spp. strains from Hospital Eugenio Espejo–HEE, Hospital Militar–HMI, Hospital Carlos Andrade Marín–HCA, and Hospital Gustavo Dominguez—HGD. These samples were received at the National Reference Center for Antimicrobial Resistance (CRN-RAM) Dr. Leopoldo Izquieta Pérez in Quito. The origin of the samples is shown in **Supplementary Table 1**. These samples have a similar phenotype to those at the Antimicrobial Resistance Reference Center as part of the national surveillance program which were isolated in Ecuador between November 2017 and April 2018.

The Antimicrobial Resistance (AMR) surveillance, the National Antimicrobial Resistance Reference Center (CRN-RAM) has defined a list of microorganisms with antimicrobial susceptibility patterns that are included in epidemiological surveillance actions in different hospitals in Ecuador. In this case, the hospitals reported the presence of resistance to carbapenems and susceptibility to third-generation cephalosporins, suggesting that this observation could be attributed to resistance mediated by $bla_{\text{OXA}-48}$ carbapenemase. Consequently, epidemiological alert control measures were implemented, indicating the possible presence of an outbreak with a common source not identified, which was

subsequently investigated through phenotypic observation in the laboratory followed by PCR and genomic sequencing. Following the last reported case of *Raoultella ornithinolytica*, hospitals conducted active surveillance for 6 months, during and no further cases were detected.

Isolates were recovered from BD CultureSwab MAXV smears with Stuart liquid medium by plating them on BD Difco nutrient agar and MacConkey agar. Plates were incubated at 37°C for 24 h. Pure colonies were selected for bacterial identification and antimicrobial susceptibility using VITEK®2 GN cards and the VITEK® 2 Compact Microbial Detection System (BioMerieux Inc., France) according to the manufacturer's recommendations.

2.2. Genome sequencing and assembly

Genomic DNA was extracted and purified using the Wizard Genomic Purification Kit (Promega). DNA extracts were used to perform WGS at the Malbran Institute in Argentina by the Illumina Miseq platform using the Nextera XT DNA library preparation kit (Illumina®, San Diego, CA, USA). Libraries were generated with a length of 300 bp paired-end reads. The quality control analysis of reads was performed using FastQC software v.011.9 (" Babraham Bioinformatics, 2022–FastQC A Quality Control Tool for High Throughput Sequence Data" n.d.). Subsequently, reads were trimmed using TrimGalore v.0.6.7 (Krueger et al., 2021) with Q > 20. The reads were assembled into contigs using SPAdes v.3.13.1 (Bankevich et al., 2012) with the default settings, and the assembly statistics were performed with QUAST v5.0.2 (Gurevich et al., 2013).

2.3. Genome identification, annotation, and phylogenomic tree construction

Bacterial identification was performed using the criterion Average Nucleotide Identity (ANI), using PyANI v.0.2.12 with default parameters (Pritchard et al., 2016). Briefly, ANI is defined as the mean nucleotide identity of orthologous gene pairs shared by two microbial genomes considered one genome >95% ANI as the same species (Barco et al., 2020). We retrieved 41 external complete genomes of the genus *Raoultella* from the NCBI portal (retrieved October 9, 2022), as well as 5 genomes of *Klebsiella* as controls for this analysis. A total of 59 genomes were included in the ANI analysis. The assembled genomes were annotated using Prokka v.1.14.6 (Seemann, 2014) and then compared for size and similarity to the reference sequence using BLAST BRIG ring imager v.0.95 (Alikhan et al., 2011).

For phylogenomic reconstruction, a set of 92 bacterial core genes was retrieved from the 59 genomes and aligned using Up-to-date Bacterial Core Gene Software (UBCG) v.3.0 (Na et al., 2018). A maximum-likelihood (ML) tree was estimated with IQ-TREE v2.0.3 (Nguyen et al., 2015) based on 1,000 bootstrap replicates. The five genomes of *Klebsiella* were used as the phylogenetic tree root.

2.4. Core and pan-genome analysis

Core and pan-genome analysis were performed using only *R. ornithinolytica* genomes, including our 13 Ecuadorian genomes and 22 complete genomes from NCBI (**Supplementary Table 1**). The 35 genomes were analyzed using Roary (Page et al., 2015), with 95% identity for blastp and a strict definition of nuclear genome (i.e., 100% of isolates with core genes) (Page et al., 2015).

TABLE 1 Primers for identification of carbapenems.

Gene	Primer	Sequence (5′-3′)	Product size (bp)
bla_{KPC}	KPC-Fm	CGTCTAGTTCTGCTGTCTTG	798
	KPC-Rm	CTTGTCATCCTTGTTAGGCG	232
bla_{NDM}	NDM-F	GGTTTGGCGATCTGGTTTTC	621
	NDM-R	CGGAATGGCTCATCACGATC	
bla_{VIM}	VIM-F	GATGGTGTTTGGTCGCATA	390
	VIM-R	CGAATGCGCAGCACCAG	
bla_{IMP}	IMP-F	GGAATAGAGTGGCTTAAYTCTC	232
	IMP-R	GGTTTAAYAAAACAACCACC	
bla_{OXA-48}	OXA48-F	GCGTGGTTAAGGATGAACAC	438
	OXA48-R	CATCAAGTTCAACCCAACCG	

2.5. Antimicrobial susceptibility test, identification, and search for antimicrobial resistance genes and plasmids using bioinformatics

Susceptibility testing was performed using a VITEK-2® AST N272 card (BioMerieux Inc., France). The modified carbapenem inactivation method (mCIM) was performed according to the Clinical and Laboratory Standards Institute (CLSI) to evaluate the presence and expression of enzymatic antimicrobial resistance mechanisms. Molecular analysis by PCR was performed to identify the following ARGs: bla_{KPC}, bla_{NDM}, bla_{VIM}, bla_{IMP}, and bla_{OXA-48}, according to previously published protocols described by Poirel et al. (2011) (**Table 1**). We used the Comprehensive Antibiotic Resistance Database (CARD) in Resistance Gene Identifier (RGI) v.5.1.0 to search ARGs (Alcock et al., 2020), and PLACNETw (Vielva et al., 2017) to search for plasmids, and circular plasmids were visualized through BRING v.0.95, in previously assembled genomes.

Incompatibility groups (Inc.) and pMLST subtypes were determined *in silico* using PlasmidFinder and pMLST (Carattoli and Hasman, 2020), and ccfind (Nishimura et al., 2017) was used to determine the circulated genome.

2.6. Gain and loss genes from ancestral reconstruction in *R. ornithinolytica*

We performed an ancestral reconstruction analysis to estimate genetic gain and loss events. We used the Wagner parsimony model with flexible gain-loss ratios for all lineages and a Poisson distribution at the root of the tree in the program COUNT (Csurös, 2010). The number of orthologs in each genome was determined using Proteinortho, with a threshold of 50% for identity and coverage (Lechner et al., 2011).

TABLE 2 Assembly statistics of *Raoultella ornithinolytica* the isolated from hospitals in Ecuador.

Code	Genebank accesión number	Sequence size	Number of contigs	GC content (%)	Median sequence size	Mean sequence size	N50 value	Number of coding sequences
HEE0686	GCF_021698415.1	5,593,786	48	55.6	6024	13822.2	281,078	5,161
HCA1847	GCF_021698795.1	5,593,768	39	55.6	3033	22461.1	400,770	5,153
HMI0464	GCF_021698545.1	5,729,712	52	55.6	17163	36475.0	349,492	5,353
HEE0460	GCF_021698375.1	5,671,979	72	55.7	28061.5	217396	290,183	5,305
HEE0459	GCF_021698805.1	5,715,189	82	55.6	5486	9560.5	195,910	5,368
HEE0406	GCF_021698245.1	5,684,786	67	55.7	10683	32022.8	220,485	5,374
HEE0315	GCF_021698635.1	5,598,932	42	55.6	4345	36781.8	738,672	5,160
HGD0386	GCF_021698595.1	5,582,782	45	55.7	398	21074.5	367,263	5,152
HMI0367	GCF_021698535.1	5,597,243	45	55.6	768	24210.1	401,817	5,105
HEE0314	GCF_021698645.1	5,597,398	43	55.6	16476.0	197527	367,263	5,160
HEE0463	GCF_021698235.1	5,596,758	97	55.6	473	57765.1	393,497	5,153
HEE0169	GCF_021698685.1	5,666,149	75	55.6	7307	22939.2	215,878	5,354
HCA1697	GCF_021699015.1	5,927,737	70	55.3	570	37620.9	363,932	5,590

3. Results

3.1. Bacterial strains

From November 2017 to April 2018, a proportional increase in *Raoultella* spp. infections were reported in tertiary hospitals in Quito and Santo Domingo de los Tsachilas, Ecuador; The strains were identified in NRLAR using the VITEK®2 GN card compact system (BioMerieux Inc., France) and confirmed as *R. ornithinolytica* by sequencing of the *rpoB* gene.

3.2. Genomic features of the *Raoultella ornithinolytica* isolated

Whole-genome sequencing sequences of the isolated *Raoultella* strains were sequenced individually and then merged. The general characteristics of the 13 assembled genomes are summarized in Table 2. The size of the genomes ranged from 5.5 Mb to 5.9 Mb. The average GC content of each genome was approximately 55.6% (Figure 1). The average number of coding sequences (CDS) in each genome was 5,485 pb, as shown in Table 2. Taxonomic identification was based on average nucleotide identity (ANI), and whole genome comparison results showed that all 13 genomes were nearly 99% concordant with the 22 *R. ornithinolytica* genome references (Figure 2). These results indicate that the isolates from different hospitals in Ecuador corresponded to *R. ornithinolytica*.

A phylogenetic tree was constructed using a set of 92 bacterial conserved genes to infer the phylogenetic relationship between our 13 genomes of *R. ornithinolytica*, and other 46 complete genomes of the genera *Raoultella* and *Klebsiella* retrieved from the NCBI portal. There are five main clades, as shown by the different colors in Figure 3. Clade I for *R. ornithinolytica*, Clade II for *R. planticola*, Clade III for *Raoultella* spp., Clade IV for *R. terrigena*, and Clade V for *Klebsiella* spp. In Clade I, nine out of thirteen Ecuadorian genomes of *R. ornithinolytica* were grouped together in a well-supported subclade, suggesting the idea of a common origin. The remaining four genomes, came from a single healthcare setting and seem to be dispersed throughout different subclades within Clade I, which suggests independent introduction events. In contrast, strains scattered in the tree may not belong to the same outbreak despite the geographical distance between the hospitals.

3.3. Pan-genome comparative analyses of *R. ornithinolytica* genomes

The pan-genome of the 35 genomes analyzed is shown in Figure 4. A total of 16,661 gene clusters (orthologs) were found, of which 3,467 genes (27.77%) were assigned to the core genome, 548 to the soft-core genome (3.29%), and 10,975 to the unique genome (65.87%) (Figure 4A). To investigate if the *R. ornithinolytica* pan-genomic is open, which it will suggest and increase in genome size due to the addition of new genomes, Heap's law was applied (Figure 4B). The trend of Heap's law diagram for the pangenome shows gradual expansion due to the addition of new genomes, with the slope continuing to increase. The curve shows the relationship between the core genome and the number of genomes, which decreased and did not exceed 5,000 genes. Thus, the pangenome of *R. ornithinolytica* is reflected in the openness of the pan-genome, which could indicate high genomic plasticity of this species and, thus, the greater potential of adaptability.

3.4. Antimicrobial resistance genes and plasmids

The results of antimicrobial susceptibility testing showed three different resistance phenotypic patterns (Figure 5 and Supplementary Table 2). The first includes resistance to ampicillin/sulbactam (SAM), piperacillin/tazobactam (TPZ),

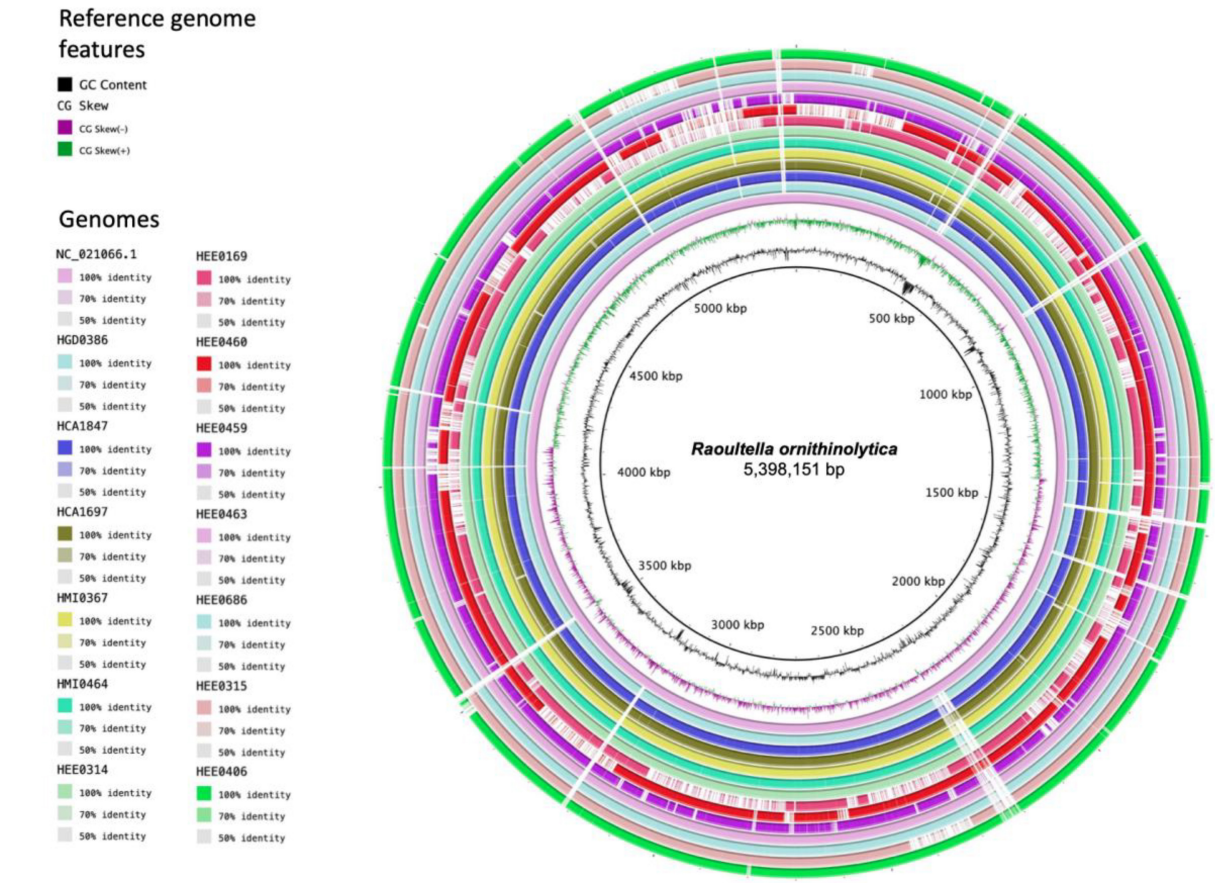


FIGURE 1
Circular genome representation of *Raoultella ornithinolytica* compared to reference genomes NC_021066.1. The genomes of the 13 strains isolated from hospitals in Ecuador are shown and the predicted coding sequences (CDS) are shown in different colors and the percentage of identity is indicated in the legend.

ertapenem (ETP), imipenem (IMP), and intermediate resistance to meropenem (MEM). The second pattern shows, resistance to SAM, TPZ, and ETP, and intermediate resistance to IMP. And the third one shows the resistance to SAM and TPZ exclusively. No specific antimicrobial resistance patterns were observed within hospitals. Therefore, the phenotypic antimicrobial resistance patterns observed in the first, second, and third groups have no consistent association with any particular hospital, as described above.

Seven *R. ornithinolytica* strains showed interesting patterns of at least two out of three carbapenems testing resistance, but appearing susceptible to third and fourth-generation cephalosporins (ceftazidime, ceftriaxone, and cefepime), which is a consistent phenotypic profile, indicating the possible presence of a $bla_{\mathrm{OXA-48}}$ carbapenemase, which was confirmed by them CIM assay and PCR (**Figure 5**).

In addition, bioinformatic analysis of antimicrobial resistance was performed using all the genomes collected using RGI software (Alcock et al., 2020; **Figure 3**). The heatmap results showed bit scores using BLAST with RGI values Perfect (green), Strict (blue), and Losse (gray) (**Figure 3B**). Among the genes predicted with a perfect bitscore with RGI analysis was the ORN-1 gene, which encoded for an Ambler class A beta-lactamase conferring resistance to penicillins and cephalosporins (Walckenaer et al., 2008). ORN-1 gene was found in all of our isolates and in 14 of the *R. ornithinolytica* genomes retrieved from NCBI for comparison purposes. The Ecuadorian genomes grouped in Clade I (**Figure 3A**) positive for ORN-1, also carried the *mphA* gene, which confers resistance to macrolides (Li et al., 2022) and the $bla_{\mathrm{OXA-48}}$ gene, which confers resistance to carbapenems; however, only seven of these isolates expressed a phenotypic resistance to carbapenems.

In addition, genes conferring resistance to aminoglycosides, beta-lactams, sulfates, tetracyclines, and dihydrofolate, such as *acrB*, *acrD*, *mdtC*, *marA*, *mdtB*, and *adeF* were present in all the Ecuadorian isolates. Nevertheless, the diversity of ARGs showed among the Ecuadorian isolates was lower when compared with the ARGs diversity from other *R. ornithinolytica* strains (**Figure 3**). Only one Ecuadorian strain (HEE0406) harbored additional antimicrobial resistance genes (*SHV1.2*, *tetA*, and *fosA*).

We have identified genes associated with antibiotic resistance in 58 isolates, and the gene *fosA* was also found in 56 isolates.

Finally, plasmid analyses revealed that the Ecuadorian *R. ornithinolytica* strains contained segments of plasmids. The partial plasmid containing the $bla_{\mathrm{OXA-48}}$ gene (22 Kb), is found in HEE0315, HEE0686, HEE0463, HEE0314, and HCA1847 strains (**Figure 6**). Unfortunately, the complete plasmid could not be

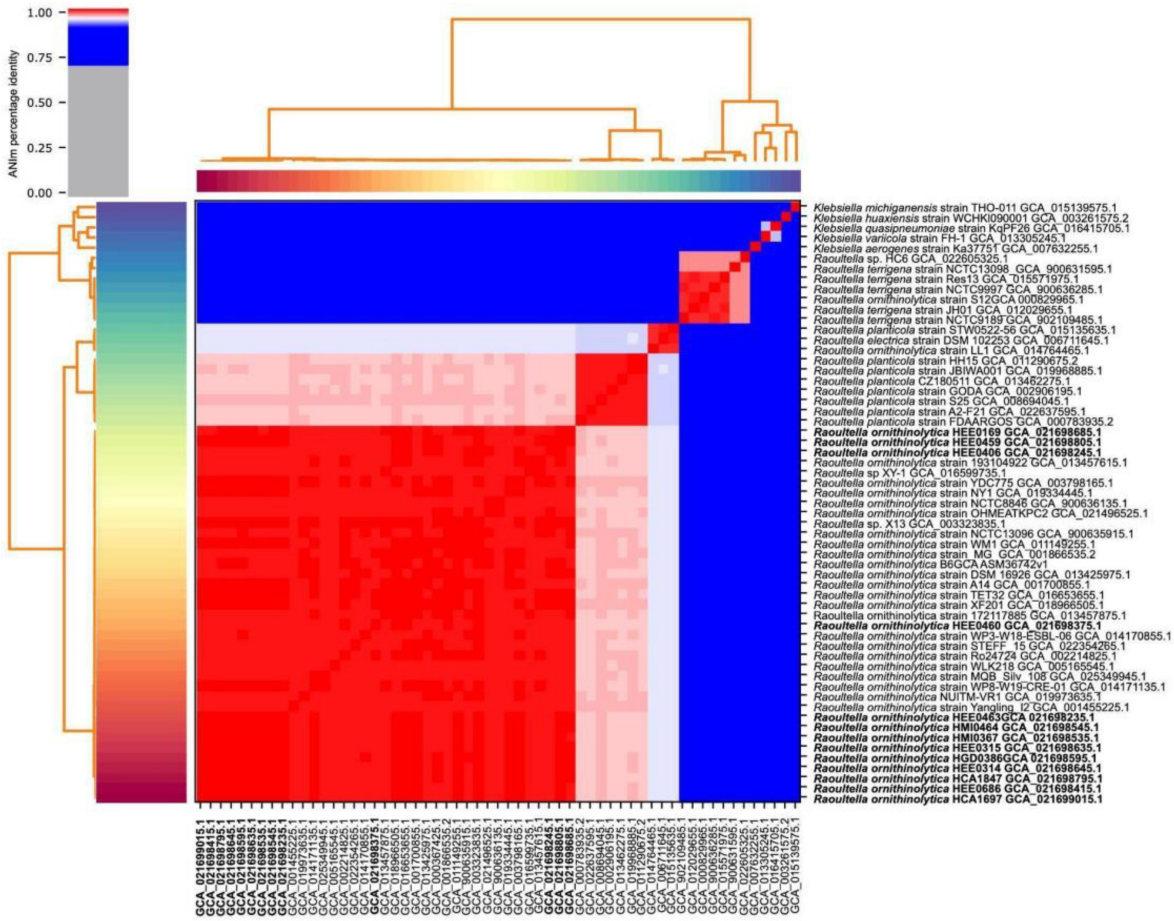


FIGURE 2
Heatmap of average nucleotide identity (ANI) comparing the 54 genomes of the genus *Raoultella* and 5 of the genus *Klebsiella* used as controls. The heatmap is the product of the matrix generated by pyANI. The red colored cells of the heatmap have >95% sequence similarity, while the blue cells have <95% similarity; once the nucleotide identity reaches 95%, the cells are colored white.

obtained, but based on previous literature and based on our knowledge, we believe that this gene could be responsible for the resistance in *R. ornithinolytica*.

3.5. Gain and loss of genes from ancestral reconstruction in *R. ornithinolytica*

Since our results revealed that the pangenome of *R. ornithinolytica* is open, the gain or loss of genes during their evolutionary history is expected. To this end, we performed an ancestral reconstruction analysis using WGS (**Figure 7**). The genomes examined shared over 4,000 protein families with their common ancestor *Klebsiella/Raoultella* (KRCA), and the reconstruction analysis predicted more gain than loss events.

Our analysis revealed that six protein families associated with antibiotic resistance were acquired from five predicted common ancestors (**Figure 7**). We hypothesize that the most remote ancestor, KRCA, harbor genes that encode an efflux pump providing resistance to fluoroquinolones (Poole, 2000), and that these genes could have been transmitted as orthologs to the analyzed genomes, except for the *K. huaxiensis* WCHKl090001 which lost that protein family. Our second hypothesis is that the common ancestor of *Raoultella* (RCA) acquired genes associated with β-lactamase resistance ORN-1/PLA-1, which were not subsequently lost. The common ancestor of *R. terrigena* (RTCA) acquired genes conferring β-lactamase resistance TER −2/DHA −1 and passed vertically to strains in this group.

In addition, the common ancestor of *R. ornithinolytica* WP8-W19-CRE-01 and *R. ornithinolytica* Ro24724 acquired genes for a pentapeptide repeat protein and aminoglycoside acetyltransferases related to fluoroquinolone and aminoglycoside resistance, respectively (Vetting et al., 2006; Reeves et al., 2020). Even though they have been isolated from different sources in China. Strain WP8-W19-CRE-01 was obtained from wastewater, whereas strain Ro24724 was obtained from necrotic tissue.

The last common ancestor with resistance protein families was the ancestor of clinical strains HEE0686, HGD0386, HCA1847, HEE0314, HMI0367, HCA1697, HEE0315, HMI0464, and HEE0463. This ancestor acquired two protein families related to the multidrug efflux pump *Tap*, and a hypothetical protein related to macrolide resistance. This ancestor may also have gained OXA-48, which was only present in this clade (**Figure 3**). In

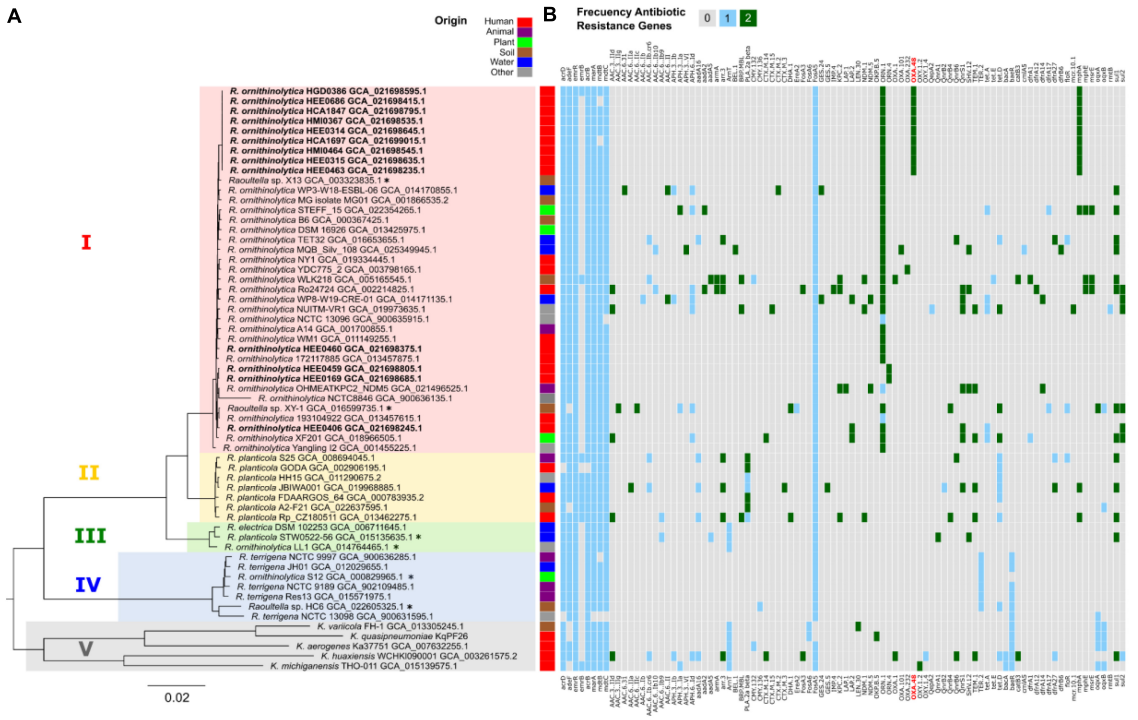


FIGURE 3
Phylogeny and resistance genes. **(A)** Phylogenetic tree of the *Raoultella* genus based on the 92 individual genes analyzed with 92 genes from UBCG and maximum likelihood with IQtree. Clades are shown, clade I in red corresponds to *R. ornithinolytica*, clade II in yellow corresponds to *R. planticola*, clade III in green corresponds to *Raoultella* spp., clade IV in blue corresponds to *R. terrigena*, and clade V in brown corresponds to *Klebsiella* spp. Asterisks indicate genomes with failed taxonomy checks in the NCBI portal. The *R. ornithinolytica* genomes in this study are shown in bold, with a clade in the upper part of the tree that groups nine genomes from Hospital Carlos Andrade Marín, Hospital Militar, Hospital Eugenio Espejo, and Hospital Gustavo Dominguez. The origin of each strain is shown in different colors. **(B)** Heatmap of resistance genes in *Raoultella* genus using Resistance Gene Identifier (RGI). The *R. ornithinolytica* genomes in this study show a pattern of resistance to bla_{OXA-48} (in red), *blaOKP-5*, and *mphA* gene, whereas the other genomes within the genus show no resistance to bla_{OXA-48}.

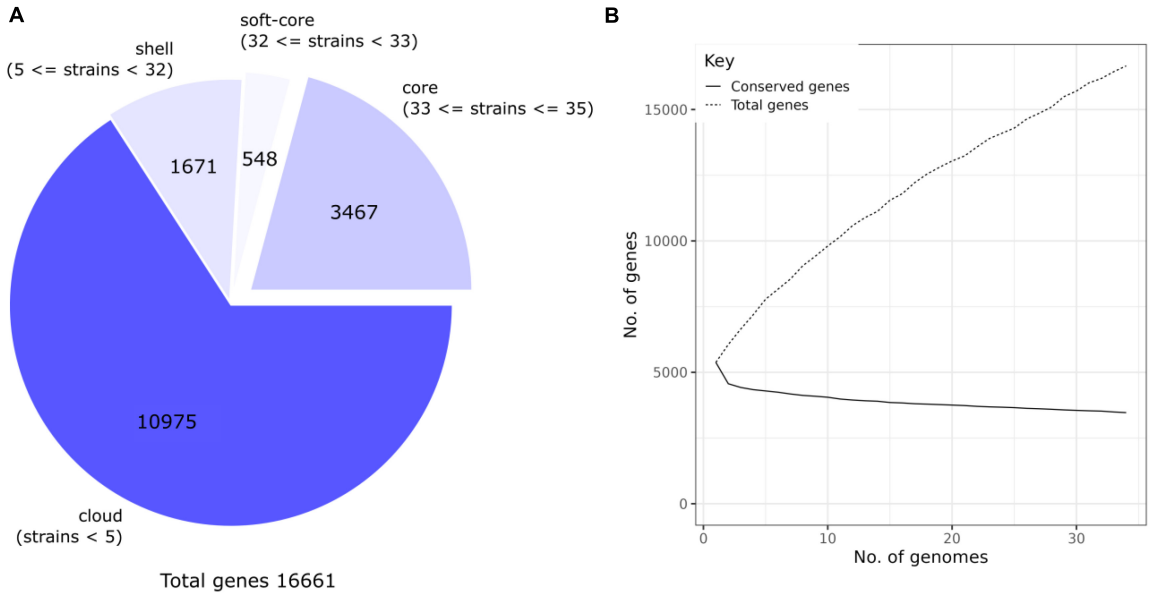


FIGURE 4
Core and pan-genome analyses of *Raoultella ornithinolytica*. **(A)** Pan-genome pie chart created with the Roary software. The gene content in the nucleus, soft nucleus, shell, and cloud is represented by the pie. The cloud contains the most genes, while the core genome contains the fewest genes. **(B)** The Heap's Law diagram is plotted. The diagram shows the conserved genes relative to the total number of genes in 35 *R. ornithinolytica* genomes. The pan-openness of the genomes reflects the diversity of the gene pool within the *R. ornithinolytica* genomes.

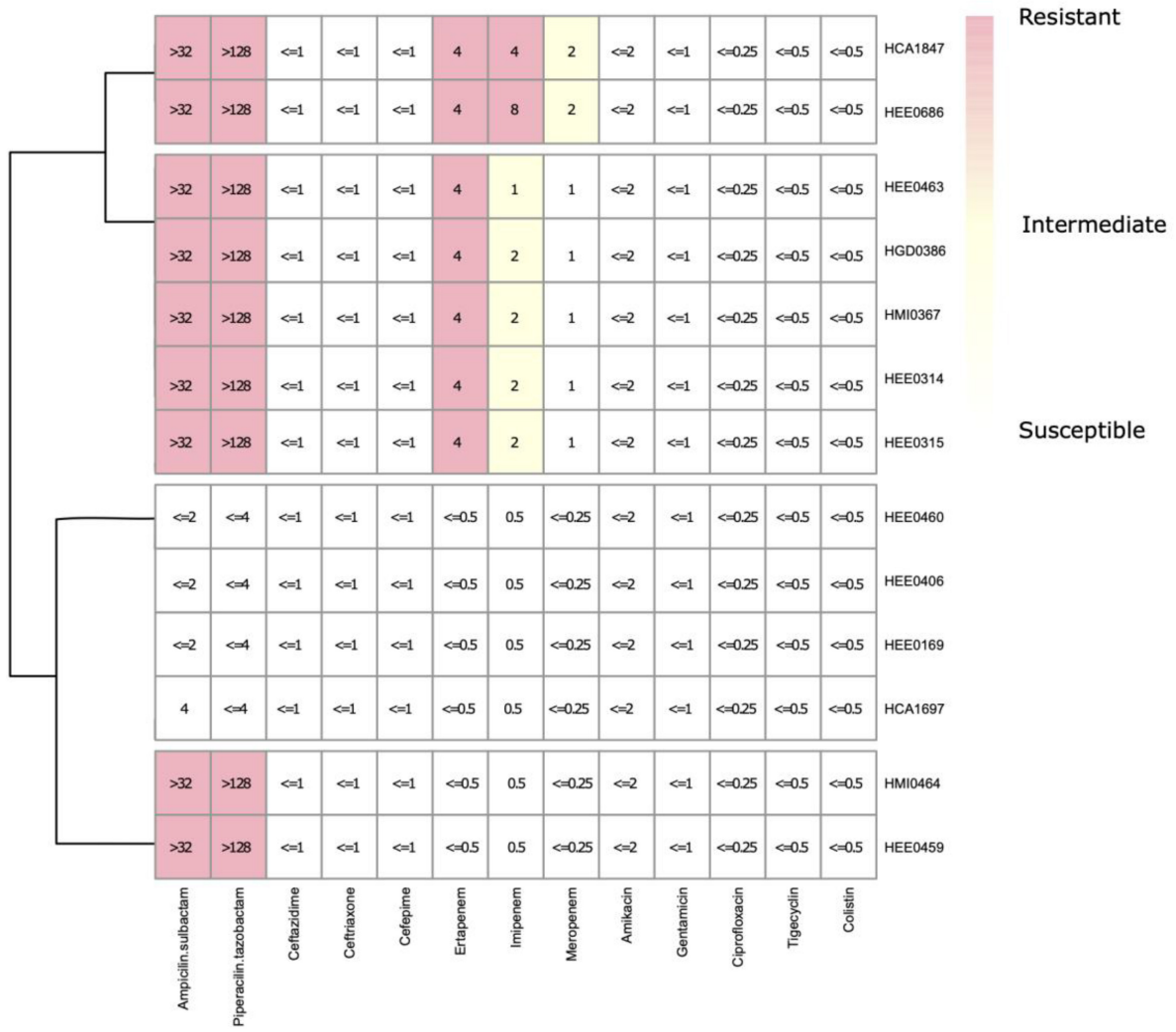


FIGURE 5
Experimental minimum inhibitory concentrations (MICs) distribution of the isolates of *Raoultella ornithinolytica* from Ecuador.

general, there are other gains/losses occurring in the genomes, but these do not present traceable ancestors in the tree because changes may have occurred recently.

4. Discussion

Raoultella ornithinolytica is a gram-negative bacterium isolated from various environments, including water and soil, as well as from various animals, including birds, humans, and other mammals (Seng et al., 2016). *R. ornithinolytica* has been shown to cause human diseases; however, its incidence may be underestimated because of the high misidentification rate of automated systems based on phenotypic characteristics (Sękowska et al., 2018). The *Raoultella* genus has been identified as *Klebsiella oxytoca* (Walckenaer et al., 2008; Park et al., 2011; Sękowska et al., 2018).

To overcome these limitations, Whole Genome Sequencing (WGS) has been used to identify *R. ornithinolytica* with greater accuracy, allowing the identification of genes that may be important for bacterial metabolism and pathogenicity (Matar et al., 2020). In Ecuador, WGS was used to identify *Raoultella* isolates, and the results showed that 13 bacteria isolated were *R. ornithinolytica*. The ANI, and taxonomic position resulted in a phylogenetic tree analysis with 92 marker genes, which showed that the strains belonged to *R. ornithinolytica* (Matar et al., 2020).

In the phylogenetic tree, nine strains from Ecuador formed a cluster, suggesting that these strains had greater genomic and phenotypic similarities than the other four strains scattered independently in the tree. This suggests that the first nine strains belonged to a single outbreak. The genomic characteristics of the 13 isolated genomes found in Ecuadorian hospitals are summarized in **Table 2**. The genomes reported in this study had a 1% size difference average variation among themselves, suggesting lost and gain events.

The phylogenetic clustering of strains within the observed phylogenetic tree indicates a potential localized transmission event or a shared exposure to a common source of infection, despite being identified in different hospitals. The clustering pattern strongly suggests the likelihood of an outbreak, which appears

10.3389/fmicb.2023.1216008

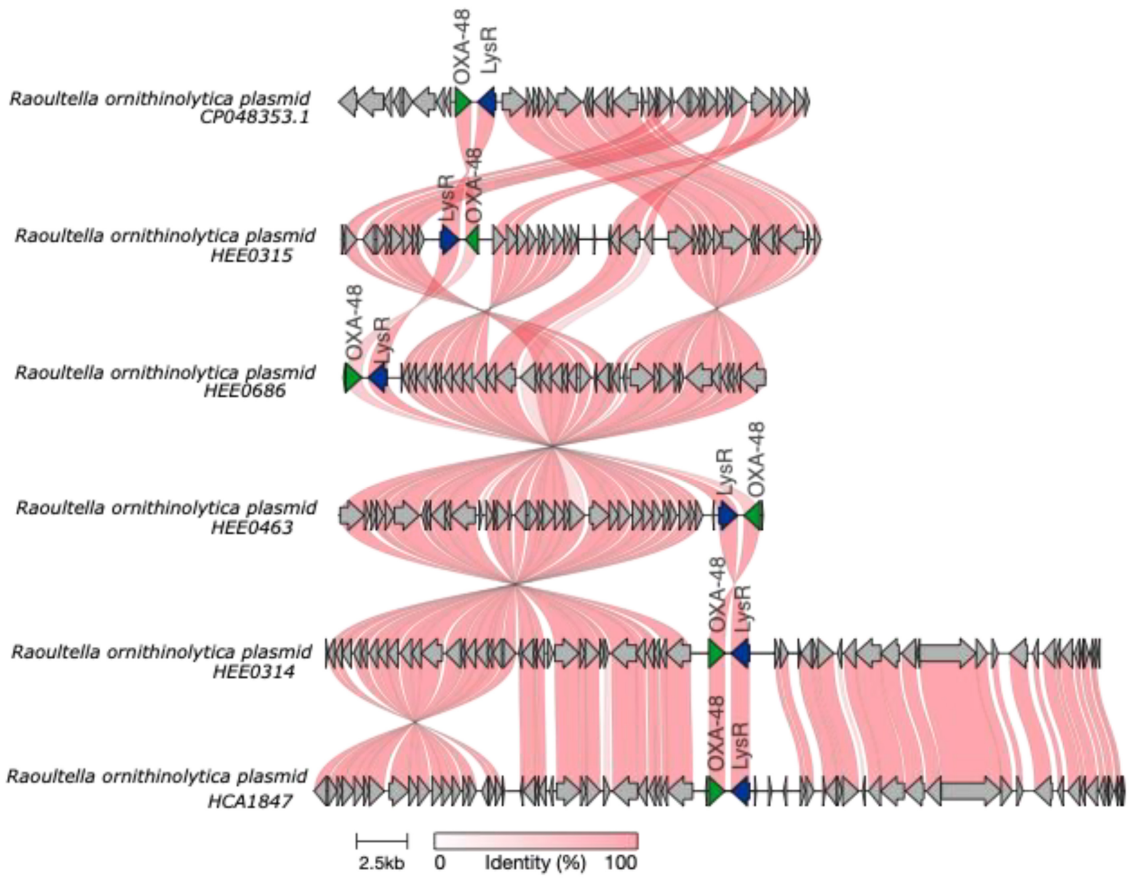


FIGURE 6
Schematic representation was made using clinker (Gilchrist and Chooi, 2021) of bla^{OXA-48} plasmids in *Raoultella ornithinolytica* genomes obtained from the clinical strains isolated in Ecuador. The frames for the genes are shown as arrows; in green is the bla^{OXA-48} gene, flagged by the *lysR* gene.

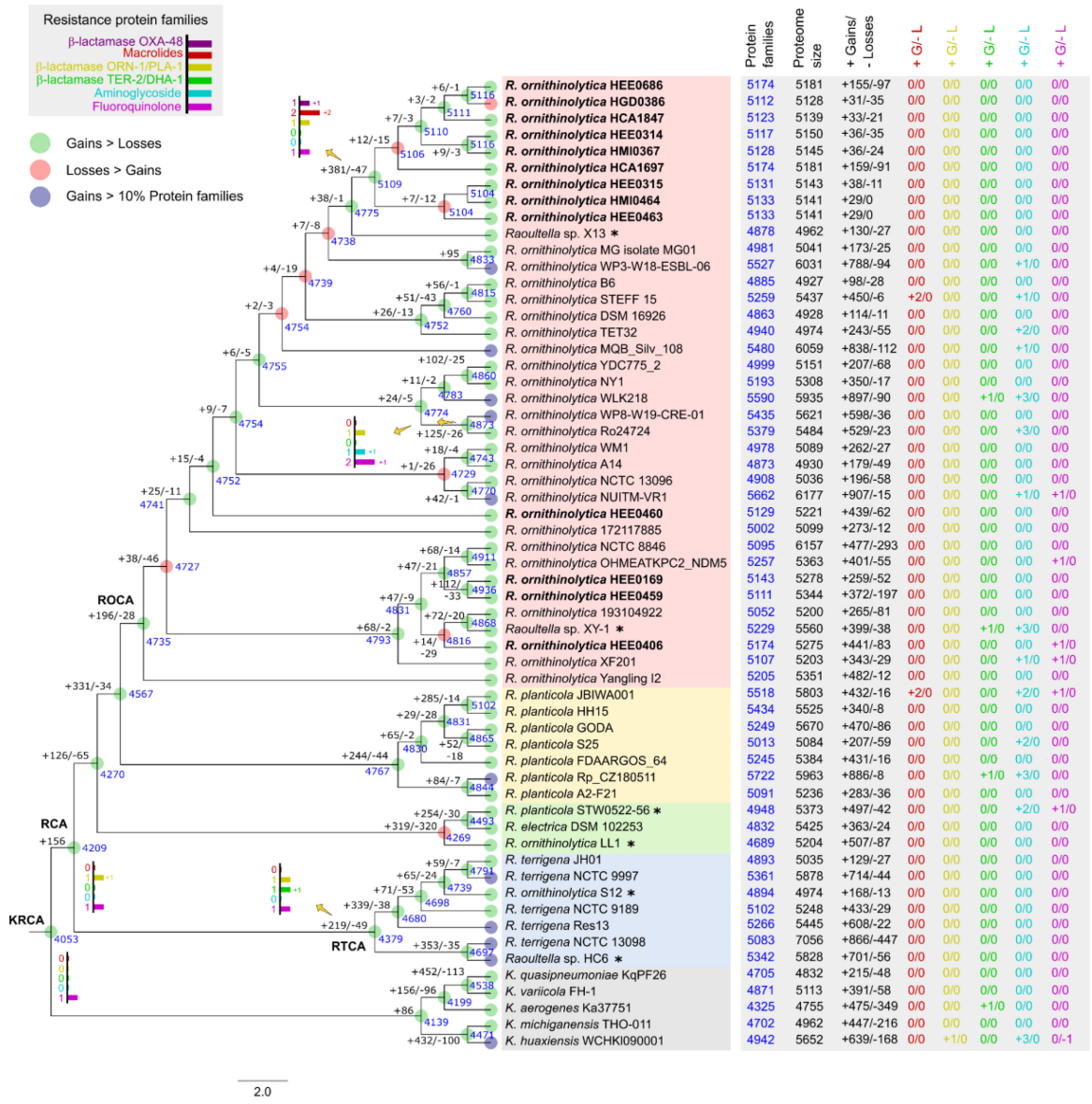
to have occurred within a specific time frame spanning from November 2017 to April 2018. Following this period, an active surveillance program was implemented for an additional 6 months, during which screening efforts were undertaken to identify strains exhibiting the same phenotypic resistance pattern for subsequent genomic sequencing. However, no strains with the identified resistance pattern were detected beyond this surveillance period.

On the other hand, the core genome was found to include 33% of the pangenome of all the *R. ornithinolytica* assemblies, using a high identity value (≥95% identity). In other species, small pan-genome size indicates a process of speciation/adaptation (Zhang et al., 2018). As more genomes were examined, the total number of genes observed in the analysis also increased, as shown in **Figure 4B**. However, this pan-genomic feature allows different strains to survive in various environmental niches (Medini et al., 2005; McInerney et al., 2017). The core genome of *R. ornithinolytica* is small compared to close bacteria, such as *Streptococcus. pneumoniae* (55.06%), *Staphylococcus aureus* subsp. *aureus* (75.23%), *Escherichia coli* (53.34%), *Salmonella enterica* subsp. *enterica* (71.23%), and *Acinetobacter baumannii* (61.36%) (Park et al., 2019).

Furthermore, results of antibiotic resistance detected by PCR showed seven positive isolates for bla_{OXA-48} and one for bla_{KPC-2}, while bioinformatic analysis revealed nine genomes from Ecuador. Carbapenem resistance in aquatic environments is caused by bla_{NDM}, bla_{OXA-48}, and bla_{KPC-2} in *R. ornithinolytica* (Hajjar et al., 2020). The emergence of carbapenem-resistant *K. pneumoniae* strains in Ecuador have increased recently. Indeed, OXA-48 carbapenemase has been detected in *K. pneumoniae* isolates from clinical samples (Reyes et al., 2020). In another case, a genome of *R. ornithinolytica* from Ecuador was reported to harbor the $bla_{OXA-48-like}$ gene (Reyes et al., 2020). Given the potential for multidrug resistance and the ability of *R. ornithinolytica* to survive in different environmental niches, it is important to continue surveillance to better understand its potential to spread antibiotic resistance.

We conducted ancestral reconstruction on all the genes in the *Raoultella* genome and identified genes that are unique to this species. In particular, we paid close attention to resistance genes to determine if they were present in the ancestral lineage. However, our analysis provides evidence to suggest that some resistance genes were ancestral, but many others are relatively recent in origin. An instance of this is the OXA-48-β-lactamase family, ORN-1/PLA-1, and fluoroquinolone resistance genes found in *Raoultella* genomes, which are likely acquired through horizontal gene transfer and not ancestral. This is consistent with the open pangenome of *R. ornithinolytica*. Ancestral reconstruction analysis reveals that each species in the genus has gained genes recently, potentially via gene transfer, indicating a worrying trend of rapid acquisition and dissemination of antibiotic resistance genes.

10.3389/fmicb.2023.1216008



	Protein families	Proteome size	+ Gains/ - Losses	+ G/- L	+ G/- L	+ G/- L	+ G/- L	+ G/- L
***R. ornithinolytica* HEE0686**	5174	5181	+155/-97	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HGD0386**	5112	5128	+31/-35	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HCA1847**	5123	5139	+33/-21	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HEE0314**	5117	5150	+36/-35	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HMI0367**	5128	5145	+36/-24	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HCA1697**	5174	5181	+159/-91	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HEE0315**	5131	5143	+38/-11	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HMI0464**	5133	5141	+29/0	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HEE0463**	5133	5141	+29/0	0/0	0/0	0/0	0/0	0/0
Raoultella sp. X13 *	4878	4962	+130/-27	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica MG isolate MG01	4981	5041	+173/-25	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica WP3-W18-ESBL-06	5527	6031	+788/-94	0/0	0/0	0/0	+1/0	0/0
R. ornithinolytica B6	4885	4927	+98/-28	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica STEFF 15	5259	5437	+450/-6	+2/0	0/0	0/0	+1/0	0/0
R. ornithinolytica DSM 16926	4863	4928	+114/-11	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica TET32	4940	4974	+243/-55	0/0	0/0	0/0	+2/0	0/0
R. ornithinolytica MQB_Silv_108	5480	6059	+838/-112	0/0	0/0	0/0	+1/0	0/0
R. ornithinolytica YDC775_2	4999	5151	+207/-68	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica NY1	5193	5308	+350/-17	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica WLK218	5590	5935	+897/-90	0/0	0/0	+1/0	+3/0	0/0
R. ornithinolytica WP8-W19-CRE-01	5435	5621	+598/-36	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica Ro24724	5379	5484	+529/-23	0/0	0/0	0/0	+3/0	0/0
R. ornithinolytica WM1	4978	5089	+262/-27	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica A14	4873	4930	+179/-49	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica NCTC 13096	4908	5036	+196/-58	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica NUITM-VR1	5662	6177	+907/-15	0/0	0/0	0/0	+1/0	+1/0
***R. ornithinolytica* HEE0460**	5129	5221	+439/-62	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica 172117885	5002	5099	+273/-12	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica NCTC 8846	5095	6157	+477/-293	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica OHMEATKPC2_NDM5	5257	5363	+401/-55	0/0	0/0	0/0	0/0	+1/0
***R. ornithinolytica* HEE0169**	5143	5278	+259/-52	0/0	0/0	0/0	0/0	0/0
***R. ornithinolytica* HEE0459**	5111	5344	+372/-197	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica 193104922	5052	5200	+265/-81	0/0	0/0	0/0	0/0	0/0
Raoultella sp. XY-1 *	5229	5560	+399/-38	0/0	0/0	+1/0	+3/0	0/0
***R. ornithinolytica* HEE0406**	5174	5275	+441/-83	0/0	0/0	0/0	0/0	+1/0
R. ornithinolytica XF201	5107	5203	+343/-29	0/0	0/0	0/0	+1/0	+1/0
R. ornithinolytica Yangling I2	5205	5351	+482/-12	0/0	0/0	0/0	0/0	0/0
R. planticola JBIWA001	5518	5803	+432/-16	+2/0	0/0	0/0	+2/0	+1/0
R. planticola HH15	5434	5525	+340/-8	0/0	0/0	0/0	0/0	0/0
R. planticola GODA	5249	5670	+470/-86	0/0	0/0	0/0	0/0	0/0
R. planticola S25	5013	5084	+207/-59	0/0	0/0	0/0	+2/0	0/0
R. planticola FDAARGOS_64	5245	5384	+431/-16	0/0	0/0	0/0	0/0	0/0
R. planticola Rp_CZ180511	5722	5963	+886/-8	0/0	0/0	+1/0	+3/0	0/0
R. planticola A2-F21	5091	5236	+283/-36	0/0	0/0	0/0	0/0	0/0
R. planticola STW0522-56 *	4948	5373	+497/-42	0/0	0/0	0/0	+2/0	+1/0
R. electrica DSM 102253	4832	5425	+363/-24	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica LL1 *	4689	5204	+507/-87	0/0	0/0	0/0	0/0	0/0
R. terrigena JH01	4893	5035	+129/-27	0/0	0/0	0/0	0/0	0/0
R. terrigena NCTC 9997	5361	5878	+714/-44	0/0	0/0	0/0	0/0	0/0
R. ornithinolytica S12 *	4894	4974	+168/-13	0/0	0/0	0/0	0/0	0/0
R. terrigena NCTC 9189	5102	5248	+433/-29	0/0	0/0	0/0	0/0	0/0
R. terrigena Res13	5266	5445	+608/-22	0/0	0/0	0/0	0/0	0/0
R. terrigena NCTC 13098	5083	7056	+866/-447	0/0	0/0	0/0	0/0	0/0
Raoultella sp. HC6 *	5342	5828	+701/-56	0/0	0/0	0/0	0/0	0/0
K. quasipneumoniae KqPF26	4705	4832	+215/-48	0/0	0/0	0/0	0/0	0/0
K. variicola FH-1	4871	5113	+391/-58	0/0	0/0	0/0	0/0	0/0
K. aerogenes Ka37751	4325	4755	+475/-349	0/0	0/0	+1/0	0/0	0/0
K. michiganensis THO-011	4702	4962	+447/-216	0/0	0/0	0/0	0/0	0/0
K. huaxiensis WCHKl090001	4942	5652	+639/-168	0/0	+1/0	0/0	+3/0	0/-1

FIGURE 7
Ancestral reconstruction of genomes of *Raoultella ornithinolytica* using COUNT software. The number of gene family gains (" + ") and losses ("−") were calculated using Wagner's parsimony are presented in each branch, and the number of gene families are shown in blue in the corresponding branch. The tree topology is based on the maximum likelihood in **Figure 3**. The bar charts show the numbers of gained/losses of the protein families related to resistance observed in common ancestors. Nodes with more gain events are indicated in green circles; nodes with more loss events are in red circles; nodes with significant genome expansion (>10% gains) are indicated in blue circles. The table shows number features of the selected genomes: protein families, proteome size, total gains/losses events, and gains/losses events of the protein families related to resistance observed in each terminal node. KRCA, *Klebsiella/Raoultella* common ancestor; RCA, *Raoultella* common ancestor; RTCA, *R. terrigena* common ancestor; ROCA, *R. ornithinolytica* common ancestor.

A gene harbored by the common ancestor of the *R. ornithinolytica* hospital clade, important for antibiotic resistance, was the IS5 family transposase IS4811, also found in some *Klebsiella* and *R. terrigena* assemblies. This transposase is important for the transfer of genetic material between bacteria, and it can contribute to the spread of antibiotic resistance genes among bacterial populations (Vandecraen et al., 2017).

Also, this ancestor contains the Multidrug efflux pump *Tap*, and a gene encoding for a hypothetical protein related to resistance to macrolides, suggesting that these genes are important for antibiotic resistance in this clade of bacteria. Multidrug efflux pump *Tap* and the hypothetical protein genes were also present in other genomes, such as *R. ornithinolytica* STEFF_15, *R. planticola* JBIWA001, and *R. planticola* Rp_CZ180511, suggesting that they may have been horizontally transferred between different *Raoultella* species.

The presence of the β-lactamase ORN-1/PLA-1 protein family since the *Raoultella* common ancestor, and its acquisition by *K. huaxiensis* WCHKl090001, suggests that this gene family is important for the survival of these bacteria in the presence of β-lactam antibiotics.

The β-lactamase TER-2, a β-lactamase DHA-1, is common to all *R. terrigena*, and according to **Figure 7**, it has orthologs in *Raoultella* sp. XY-1, *R. planticola* Rp_CZ180511 and *K. aerogenes* Ka37751. These genes likely provide these bacteria with resistance to β-lactam antibiotics, which are commonly used to treat bacterial infections. The acquisition of these resistance genes may have facilitated the spread of antibiotic resistance among different bacterial species, which is a major public health concern (Pérez-Llarena et al., 2014).

The only common ancestor with an aminoglycoside resistance family was found between the *R. ornithinolytica* WP8-W19-CRE-01 and *R. ornithinolytica* Ro24724 genomes, including the gene for Aminoglycoside N(6′)-acetyltransferase type 1. The other aminoglycoside resistance gene families (*aac*, *aad*, *arm*) were gained by the strains.

The common ancestor for *Raoultella* and *Klebsiella* has one fluoroquinolone resistance protein family, encoding for Efflux pump periplasmic linker *BepF*, related to this activity (Li and Nikaido, 2009). All the reported *Raoultella* genomes from Ecuador keep these genes. The *Raoultella sp*. XY-1 gained two copies of this gene and, like some other later characters, acquired the repeat protein Qnr (**Figure 3**).

In conclusion, continued monitoring and research of *R. ornithinolytica* are necessary to better understand its potential for multidrug resistance. This study highlights the importance of using Whole Genome Sequencing as a valuable tool in identifying *R. ornithinolytica* with greater accuracy and overcoming the limitations of automated systems based on phenotypic characteristics. Finally, the study emphasizes the emergence of multidrug-resistant enterobacteria, including *R. ornithinolytica*, which raises concerns regarding the spread of antibiotic resistance and the potential for horizontal gene transfer.

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

JV, HC-S, JR-V, and MCG: conceptualization. HC-S, JR-V, VA, UR-C, PM-R, FV, and MCG: methodology. JV, HC-S, MG, JR-V, and VA: writing–original draft preparation. JV, HC-S, JR-V, UR-C, VA, MG, SD-R, and MCG: writing–review and editing. HC-S, JR-V, UR-C, and VA: formal analysis. JV, HC-S, JR-V, SD-R, MCG, and MG: supervision and review. JV and MCG: project administration and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by the Instituto Nacional de Investigación en Salud Pública "Dr. Leopoldo Izquieta Pérez," Quito, Ecuador and the Louisiana Board of Regents LEQSF(2022-25)-RD-A-33.

Acknowledgments

We thank Juan Manuel Hurtado Ramírez for their support in the installation of bioinformatics programs. We also thank the Instituto de Biotecnología UNAM for giving us access to its computer cluster.

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.2023.1216008/full#supplementary-material

References

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

Alikhan, N., Petty, N., Ben Zakour, N., and Beatson, S. (2011). BLAST ring image generator (BRIG): Simple prokaryote genome comparisons. *BMC Genom.* 12:402. doi: 10.1186/1471-2164-12-402

Alves, M., Dias, R., de Castro, A., Riley, L., and Moreira, B. (2006). Identification of clinical isolates of indole-positive and indole-negative *Klebsiella* spp. *J. Clin. Microbiol.* 44, 3640–3646. doi: 10.1128/JCM.00940-06

Avellaneda, H., Arbeli, Z., Teran, W., and Roldan, F. (2020). Transformation of TNT, 2,4-DNT, and PETN by *Raoultella planticola* M30b and Rhizobium radiobacter M109 and exploration of the associated enzymes. *World J. Microbiol. Biotechnol.* 36:190. doi: 10.1007/s11274-020-02962-8

Babraham Bioinformatics (2022). *FastQC A quality control tool for high throughput sequence data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Oct 13, 2022).

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A., 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

Barco, R., Garrity, G., Scott, J., Amend, J., Nealson, K., and Emerson, D. A. (2020). Genus definition for bacteria and archaea based on a standard genome relatedness index. *mBio* 11:e02475–19. doi: 10.1128/mBio.02475-19

Beye, M., Hasni, I., Seng, P., Michelle, C., La Scola, B., Raoult, D., et al. (2018). Genomic analysis of a *Raoultella ornithinolytica* strain causing prosthetic joint infection in an immunocompetent patient. *Sci. Rep.* 8:9462. doi: 10.1038/s41598-018-27833-z

Boye, K., and Hansen, D. (2003). Sequencing of 16S rDNA of *Klebsiella*: Taxonomic relations within the genus and to other *Enterobacteriaceae*. *Int. J. Med. Microbiol.* 292, 495–503. doi: 10.1078/1438-4221-00228

Bravo, F. (2006). "Faculty opinions recommendation of sequencing of 16S rDNA of klebsiella: Taxonomic relations within the genus and to other *Enterobacteriaceae*," in *Faculty opinions – post-publication peer review of the biomedical literature*. doi: 10.3410/f.1044055.493961

Carattoli, A., and Hasman, H. (2020). PlasmidFinder and in silico pMLST: Identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.* 2075, 285–294. doi: 10.1007/978-1-4939-9877-7_20

Cleveland, K., Mazumder, S., and Gelfand, M. (2014). Association of *Raoultella bacteremia* with diseases of the biliary tract. *Scand. J. Infect. Dis.* 46, 541–542. doi: 10.3109/00365548.2014.896032

Csurös, M. (2010). Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912. doi: 10.1093/bioinformatics/btq315

Drancourt, M., Bollet, C., Carta, A., and Rousselier, P. (2001). Phylogenetic analyses of *Klebsiella* species delineate *Klebsiella* and *Raoultella* gen. nov., with description of *Raoultella ornithinolytica* comb. nov., *Raoultella terrigena* comb. nov. and *Raoultella planticola* comb. nov. *Int. J. Syst. Evol. Microbiol.* 51, 925–932. doi: 10.1099/00207713-51-3-925

Galata, V., Backes, C., Laczny, C., Hemmrich-Stanisak, G., Li, H., Smoot, L., et al. (2018). Comparing genome versus proteome-based identification of clinical bacterial isolates. *Brief Bioinform.* 19, 495–505. doi: 10.1093/bib/bbw122

García-Lozano, T., Pascual Plá, F., and Aznar Oroval, E. (2013). *Raoultella ornithinolytica* in urinary tract infections. Clinical and microbiological study of a series of 4 oncologic patients. *Med. Clin.* 141, 138–139. doi: 10.1016/j.medcli.2012.11.021

Gilchrist, C., and Chooi, Y. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* 37, 2473–2475. doi: 10.1093/bioinformatics/btab007

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

Hajjar, R., Ambaraghassi, G., Sebajang, H., Schwenter, F., and Su, S. (2020). *Raoultella ornithinolytica*: Emergence and resistance. *Infect Drug Resist.* 13, 1091–1104. doi: 10.2147/IDR.S191387

Haruki, Y., Hagiya, H., Sakuma, A., Murase, T., Sugiyama, T., and Kondo, S. (2014). Clinical characteristics of *Raoultella ornithinolytica* bacteremia: A case series and literature review. *J. Infect Chemother.* 20, 589–591. doi: 10.1016/j.jiac.2014.05.005

Iovleva, A., Mettus, R. T., McElheny, C. L., Griffith, M. P., Mustapha, M. M., Pasculle, A. W., et al. (2019). High-level carbapenem resistance in OXA-232-producing *Raoultella ornithinolytica* triggered by ertapenem therapy. *Antimicrob. Agents Chemother.* 64:e1335-19. doi: 10.1128/AAC.01335-19

Kanki, M., Yoda, T., Tsukamoto, T., and Shibata, T. (2002). *Klebsiella pneumoniae* produces no histamine: *Raoultella planticola* and *Raoultella ornithinolytica* strains are histamine producers. *Appl. Environ. Microbiol.* 68, 3462–3466. doi: 10.1128/AEM.68.7.3462-3466.2002

Krueger, F., James, F., Ewels, P., Afyounian, E., and Schuster-Boeckler, B. (2021). *FelixKrueger/TrimGalore: V0.6.7*. Hawaii: Zenodo, doi: 10.5281/zenodo.5127899

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P., and Prohaska, S. (2011). Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinform.* 12:124. doi: 10.1186/1471-2105-12-124

Li, X.-Z., and Nikaido, H. (2009). Efflux-mediated drug resistance in bacteria. *Drugs* 69, 1555–1623.

Li, Y., Qiu, Y., Gao, Y., Chen, W., Li, C., Dai, X., et al. (2022). Genetic and virulence characteristics of a *Raoultella planticola* isolate resistant to carbapenem and tigecycline. *Sci. Rep.* 12:3858. doi: 10.1038/s41598-022-07778-0

Martínez, J., Martínez, L., Rosenblueth, M., Silva, J., and Martínez-Romero, E. (2004). How are gene sequence analyses modifying bacterial taxonomy? The case of *Klebsiella*. *Int. Microbiol.* 7, 261–268.

Matar, G. M., Andremont, A., and Bazzi, W. (2020). Editorial: Combating antimicrobial resistance – a one health approach. *Front. Cell. Infect. Microbiol.* 9:458. doi: 10.3389/fcimb.2019.00458

Mau, N., and Ross, L. (2010). *Raoultella ornithinolytica* bacteremia in an infant with visceral heterotaxy. *Pediatr. Infect. Dis. J.* 29, 477–478. doi: 10.1097/INF.0b013e3181ce9227

McInerney, J. O., McNally, A., and O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nat. Microbiol.* 2:17040.

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

Na, S., Kim, Y., Yoon, S., Ha, S., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi: 10.1007/s12275-018-8014-6

Nguyen, L., Schmidt, H., 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

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., et al. (2017). Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere* 2:e359-16. doi: 10.1128/mSphere.00359-16

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

Park, J., Hong, K., Lee, H., Choi, S., Song, S., Song, K., et al. (2011). Evaluation of three phenotypic identification systems for clinical isolates of *Raoultella ornithinolytica*. *J. Med. Microbiol.* 60, 492–499. doi: 10.1099/jmm.0.020768-0

Park, S.-C., Lee, K., Kim, Y. O., Won, S., and Chun, J. (2019). Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Front. Microbiol.* 10:834. doi: 10.3389/fmicb.2019.00834

Pérez-Llarena, F. J., Zamorano, L., Kerff, F., Beceiro, A., García, P., Miró, E., et al. (2014). Genetic and kinetic characterization of the novel AmpCβ-lactamases DHA-6 and DHA-7. *Antimicrob. Agents Chemother.* 58, 6544–6549. doi: 10.1128/AAC.03144-14

Poirel, L., Walsh, T., Cuvillier, V., and Nordmann, P. (2011). Multiplex PCR for detection of acquired carbapenemase genes. *Diagn. Microbiol. Infect Dis.* 70, 119–123. doi: 10.1016/j.diagmicrobio.2010.12.002

Poole, K. (2000). Efflux-mediated resistance to fluoroquinolones in gram-negative bacteria. *Antimicrob. Agents Chemother.* 44, 2233–2241.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Analyt. Methods* 8, 12–24.

Reeves, C. M., Magallon, J., Rocha, K., Tran, T., Phan, K., Vu, P., et al. (2020). Aminoglycoside 6'-N-acetyltransferase type Ib [AAC(6')-Ib]-mediated aminoglycoside resistance: Phenotypic conversion to susceptibility by silver ions. *Antibiotics (Basel)* 10:29. doi: 10.3390/antibiotics10010029

Reyes, J., Villavicencio, F., Villacís, J., Pavón, E., Campoverde, N., Espinel, M., et al. (2020). First report of a clinical isolate of blaOXA-48- carbapenemase producing *Raoultella ornithinolytica* in South America. *Rev. Argent Microbiol.* 52, 82–83. doi: 10.1016/j.ram.2019.02.002

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

Sękowska, A., Mikucka, A., and Gospodarek-Komkowska, E. (2018). Identification of *Raoultella* spp.: Comparison of three methods. *Indian J. Med. Microbiol.* 36, 197–200. doi: 10.4103/ijmm.IJMM_17_99

Seng, P., Boushab, B., Romain, F., Gouriet, F., Bruder, N., Martin, C., et al. (2016). Emerging role of *Raoultella ornithinolytica* in human infections: A series of cases and review of the literature. *Int. J. Infect. Dis.* 45, 65–71. doi: 10.1016/j.ijid.2016.02.014

Tafoukt, R., Touati, A., Leangapichart, T., Bakour, S., and Rolain, J. (2017). Characterization of OXA-48-like-producing *Enterobacteriaceae* isolated from river water in Algeria. *Water Res.* 120, 185–189. doi: 10.1016/j.watres.2017.04.073

Vandecraen, J., Chandler, M., Aertsen, A., and Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.* 43, 709–730.

Vetting, M. W., Hegde, S. S., Fajardo, J. E., Fiser, A., Roderick, S. L., Takiff, H. E., et al. (2006). Pentapeptide repeat proteins. *Biochemistry* 45, 1–10.

Vielva, L., de Toro, M., Lanza, V., and de la Cruz, F. (2017). PLACNETw: A web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics* 33, 3796–3798. doi: 10.1093/bioinformatics/btx462

Walckenaer, E., Leflon-Guibout, V., and Nicolas-Chanoine, M. (2008). How to identify *Raoultella* spp. including *R. ornithinolytica* isolates negative for ornithine decarboxylase? The reliability of the chromosomal bla gene. *J. Microbiol. Methods* 75, 405–410. doi: 10.1016/j.mimet.2008.07.011

Weng, X., Shi, Q., Wang, S., Shi, Y., Sun, D., and Yu, Y. (2020). The characterization of OXA-232 carbapenemase-producing ST437 *Klebsiella pneumoniae* in China. *Can. J. Infect. Dis. Med. Microbiol.* 2020:5626503. doi: 10.1155/2020/5626503

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

Zhang, X., Liu, X., Yang, F., and Chen, L. (2018). Pan-genome analysis links the hereditary variation of *Leptospirillum ferriphilum* with its evolutionary adaptation. *Front. Microbiol.* 9:577. doi: 10.3389/fmicb.2018.00577

Zou, H., Berglund, B., Xu, H., Chi, X., Zhao, Q., Zhou, Z., et al. (2020). Genetic characterization and virulence of a carbapenem-resistant *Raoultella ornithinolytica* isolated from well water carrying a novel megaplasmid containing blaNDM-1. *Environ. Pollut.* 260: 114041. doi: 10.1016/j.envpol.2020.114041

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 07 September 2023
DOI 10.3389/fmicb.2023.1243987

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Etienne Z. Gnimpieba,
University of South Dakota, United States
Sergey N. Gavrilov,
Russian Academy of Sciences (RAS), Russia

*CORRESPONDENCE
Huaqun Yin
✉ yinhuaqun_cs@sina.com
Delong Meng
✉ delong.meng@csu.edu.cn

RECEIVED 21 June 2023
ACCEPTED 21 August 2023
PUBLISHED 07 September 2023

CITATION
Li L, Zhou L, Jiang C, Liu Z, Meng D, Luo F, He Q and Yin H (2023) AI-driven pan-proteome analyses reveal insights into the biohydrometallurgical properties of *Acidithiobacillia*. *Front. Microbiol.* 14:1243987. doi: 10.3389/fmicb.2023.1243987

COPYRIGHT
© 2023 Li, Zhou, Jiang, Liu, Meng, Luo, He and Yin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AI-driven pan-proteome analyses reveal insights into the biohydrometallurgical properties of *Acidithiobacillia*

Liangzhi Li[1,2], Lei Zhou[3], Chengying Jiang[4,5], Zhenghua Liu[1,2], Delong Meng[1,2]*, Feng Luo[6], Qiang He[7] and Huaqun Yin[1,2]*

[1]School of Minerals Processing and Bioengineering, Central South University, Changsha, China, [2]Key Laboratory of Biometallurgy of Ministry of Education, Central South University, Changsha, China, [3]Beijing Research Institute of Chemical Engineering and Metallurgy, Beijing, China, [4]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China, [5]University of Chinese Academy of Sciences, Beijing, China, [6]School of Computing, Clemson University, Clemson, SC, United States, [7]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Knoxville, TN, United States

Microorganism-mediated biohydrometallurgy, a sustainable approach for metal recovery from ores, relies on the metabolic activity of acidophilic bacteria. *Acidithiobacillia* with sulfur/iron-oxidizing capacities are extensively studied and applied in biohydrometallurgy-related processes. However, only 14 distinct proteins from *Acidithiobacillia* have experimentally determined structures currently available. This significantly hampers in-depth investigations of *Acidithiobacillia*'s structure-based biological mechanisms pertaining to its relevant biohydrometallurgical processes. To address this issue, we employed a state-of-the-art artificial intelligence (AI)-driven approach, with a median model confidence of 0.80, to perform high-quality full-chain structure predictions on the pan-proteome (10,458 proteins) of the type strain *Acidithiobacillia*. Additionally, we conducted various case studies on *de novo* protein structural prediction, including sulfate transporter and iron oxidase, to demonstrate how accurate structure predictions and gene co-occurrence networks can contribute to the development of mechanistic insights and hypotheses regarding sulfur and iron utilization proteins. Furthermore, for the unannotated proteins that constitute 35.8% of the *Acidithiobacillia* proteome, we employed the deep-learning algorithm DeepFRI to make structure-based functional predictions. As a result, we successfully obtained gene ontology (GO) terms for 93.6% of these previously unknown proteins. This study has a significant impact on improving protein structure and function predictions, as well as developing state-of-the-art techniques for high-throughput analysis of large proteomic data.

KEYWORDS

proteome, biohydrometallurgy, protein structure prediction, gene co-occurrence network analysis, *Acidithiobacillia*

1. Introduction

Biohydrometallurgy, including bioleaching and biomining, involves the accelerated dissolution of sulfidic minerals by acidophilic chemolithotrophic microorganisms to recover metals. In biohydrometallurgy operations, pulverized copper ores are piled up, inoculated with solutions containing specific microbiota and sulfuric acid, and aerated to facilitate the microbial oxidation of iron and sulfur compounds (Valdés et al., 2008). *Acidithiobacillia,* a member of the earliest and most extensively studied microbial consortia, has been widely employed in various biohydrometallurgy processes (Banderas and Guiliani, 2013; Campodonico et al., 2016; Li et al., 2019; Inaba et al., 2020). *Acidithiobacillus* is the type genus of the order *Acidithiobacillales* (type order of the class *Acidithiobacillia*) (Parte, 2014). *Acidithiobacillus* displays the central traits (e.g., sulfur/iron oxidation, CO_2 fixation, heavy metal resistance) of the deep-branching *Proteobacteria* class *Acidithiobacillia* (Moya-Beltrán et al., 2021). The representative species of *Acidithiobacillus*, *Acidithiobacillus ferrooxidans,* is a Gram-negative, strictly acidophilic, chemolithoautotrophic bacterium that thrives optimally at temperatures around 30°C and pH levels of 1.8–2.2. It is commonly found in acidic environments such as acidified mineral drainages, coal deposits, and sulfuric springs (Moya-Beltrán et al., 2021). In our previous studies, we found that frequent horizontal gene transfer (HGT) of genes vital for survival, such as heavy metal resistance, have driven the adaptation of *Acidithiobacillia* to hostile biohydrometallurgy environments (Li et al., 2019; Zhang et al., 2019). *Acidithiobacillia* exhibit remarkable abilities in the efficient dissimilatory oxidation of various reduced inorganic sulfur compounds (RISCs) (Wang et al., 2019b) and are resistant to heavy metals (Li et al., 2019). Additionally, *Acidithiobacillia* can grow by oxidizing ferrous iron Fe(II) to ferric iron Fe(III) in acidic solutions, with oxygen serving as the terminal electron acceptor (Liu et al., 2013). These combined physiological traits (sulfur and iron oxidation, acid and metal resistance) account for the widespread commercial application of *Acidithiobacillia* in biotechnologies related to the dissolution of sulfide and metallic minerals, as well as the extraction of valuable metals (Zhang et al., 2018a). Furthermore, researchers are interested in modifying *Acidithiobacillus* to become an electrochemically active bacterium (EAB) for recycling electronic waste (Wang et al., 2009) and for biofuel production from carbon dioxide using reduced iron as the sole energy source (Guan et al., 2017).

The various characteristic abilities and other life-sustaining aspects of *Acidithiobacillia* are determined by the protein machinery it encodes and expresses (Ramírez et al., 2004; Vera et al., 2013). The sulfur oxidation pathway of *Acidithiobacillia* typically involves several steps: sulfide species are oxidized to elemental sulfur by sulfide:quinone oxidoreductase (SQR), sulfide species can be converted to sulfite through sulfite reductase (Dsr), tetrathionate is converted to sulfite via tetrathionate hydrolase (Ttr) and sulfotransferase, sulfite is reversibly oxidized to sulfate through adenylylsulfate reductase (AprA) and sulfate adenylate transferase (SAT), and sulfate is transported using a sulfate transporter. Additionally, *Acidithiobacillia* primarily relies on proteins encoded by the rus gene operon (with rusticyanin as the core protein) and a high potential iron-sulfur protein (HiPIP), encoded by Iro, for iron oxidation. The HiPIP protein acts as the primary electron acceptor from Fe(II) in an alternative electron transfer pathway (Bruscella et al., 2005; Valdés et al., 2008; Quatrini et al., 2009). Over the last few decades, researchers have resolved the structure of 14 different proteins from *Acidithiobacillia* [search of the Protein Data Bank (PDB) database (Goodsell et al., 2020) with the query keyword "*Acidithiobacillus/Acidithiobacillia*"], the majority of which are involved in metabolisms of energy substrates (e.g., sulfur compounds and iron). For instance, Botuyan et al. (1996) and Walter et al. (1996) characterized the structure of the iron oxidation protein rusticyanin from *A. ferrooxidans,* which provided insights into the mechanism of its enhanced acid stability and redox potential. This was soon followed by the structure of electron transfer protein C(4)-Cytochrome of *A. ferrooxidans*, resolved by Abergel et al. (2003) and then, Cherney et al. (2010, 2012) determined the structure of sulfide:quinone oxidoreductase and its variants from *A. ferrooxidans,* from which a novel reaction mechanism utilizing the Cys-S-S as the nucleophile to attack the cofactor was proposed. More recently, the crystal structure of tetrathionate hydrolase from *A. ferrooxidans* was resolved, which suggested a novel cysteine-independent tetrathionate hydrolysis mechanism (Kanao et al., 2021). Despite these efforts, the majority of other proteins from *A. ferrooxidans* still lack three-dimensional (3D) structures. This includes proteins directly involved in sulfur/iron utilization, such as the sulfate transporter and ferrous iron transporter, which play a crucial role in its biohydrometallurgy ability. This lack of protein structure data hinders further investigations into the molecular mechanisms of these proteins. One likely reason for this is that the experimental determination of a protein's structure remains a time-consuming and expensive process (Bill et al., 2011; Lin, 2018).

Two influential artificial intelligence (AI)-driven algorithms, AlphaFold2 (Senior et al., 2020) and RoseTTAFold (Baek et al., 2021), have demonstrated their abilities to crack the long-lasting "protein-folding challenge." Both show the strength to predict a wide range of complicated protein structures accurately and quickly using solely the amino acid sequences. *Homo sapiens* was the first species whose proteome to be extended to a structural coverage scale that encompasses its near entirety (98.5%) by employing the above-mentioned AI-based algorithm predictions (Tunyasuvunakool et al., 2021). However, myriads of other organisms including industrially important and biologically significant species like *A. ferrooxidans* are still highly underrepresented in the PDB database. Researchers of these organisms would be greatly benefited if the structures of their proteome are made available. Against this background, we choose the *Acidithiobacillia* pan-proteome (Parte, 2014) as our research subject, and expanded the structural coverage of the *Acidithiobacillia* to the entire pan-proteome (10,458 proteins) with full-chain predictions through application of the advanced AI-driven program AlphaFold2 (Senior et al., 2020) and RosettaFold (Baek et al., 2021).

The objective of this study was to predict the structure of the complete proteome (10,458 proteins) of *Acidithiobacillia*. The study also aimed to conduct case studies on the sulfur/iron utilizing processes, which are currently not well understood, and to provide raw structural data that can be further analyzed in detail. These investigations have significant scientific implications for enhancing predictions of protein structure and function, and for advancing advanced techniques for analyzing large proteomic datasets.

2. Results

2.1. Protein clustering and full-length protein structure predictions

A total of 129 available genomes (isolate) of the class *Acidithiobacillia* were obtained for protein clustering and pangenome analysis. The pangenome of the 129 *Acidithiobacillia* genomes consisted of 10,458 gene families, while the core genome contained 29 gene families (Figure 1A). Analysis of the core and pangenome revealed that the pangenome followed a power-law regression function [*Ps* (n) = 4688.87 $n^{0.36}$], indicating an "open" pangenome. On the other hand, the core genome followed an exponential regression [*Fc* (n) = 2724.57 $e^{-0.22n}$] (Figure 1A). The open pangenome suggests that *Acidithiobacillia* species may undergo gene exchange in order to enhance their functional profiles. The functional COG annotation (Figure 1B) reveals that the core genome has a higher proportion of genes classified in COG categories J (translation, ribosomal structure, and biogenesis), C (energy production and conversion), O (posttranslational modification, protein turnover, and chaperones), F (nucleotide transport and metabolism), and H (coenzyme transport and metabolism), which are associated with fundamental biological functions. On the other hand, the accessory genome and strain-specific genes are skewed toward COG categories G (carbohydrate transport and metabolism), L (replication, recombination, and repair), P (inorganic ion transport and metabolism), and N (cell motility). It is likely that these categories are linked to the adaptation of *Acidithiobacillia* to oligotrophic, metal-laden, and acidic environments, which can cause DNA damage.

We predicted structures for the pan-proteome of *Acidithiobacillia*. The resulting structural dataset covers the whole pan-proteome (10,458 proteins) with full-chain predictions. The predictions made by AlphaFold2 (Senior et al., 2020) agreed well with those made by RoseTTAFold (Baek et al., 2021), indicating reliable predictions. The comparison results show that the predicted models from both methods agreed well with each other, giving average pairwise TM-score of 0.93 and average pairwise root mean square deviations (RMSD) of 1.58. The average and median of model confidences are 0.77 and 0.80, respectively, with 69.4% (2,183/3,147) of all predicted models having a confidence over 0.75, and among these models, 44.4% (969/2,183) have a confidence over 0.85 (Figure 1C). The predicted model of ACK80295 (GNAT family acetyltransferase) had the highest confidence 0.95 (Supplementary Figure 1). Regarding COG categories, proteins assigned to COG F (nucleotide transport and metabolism), COG H (coenzyme transport and metabolism), and COG J (translation, ribosomal structure, and biogenesis) had the highest average confidences (0.86, 0.85, and 0.84, respectively) (Figure 1D). Additionally, we found that the prediction confidence was not correlated with protein sequence length (data not shown).

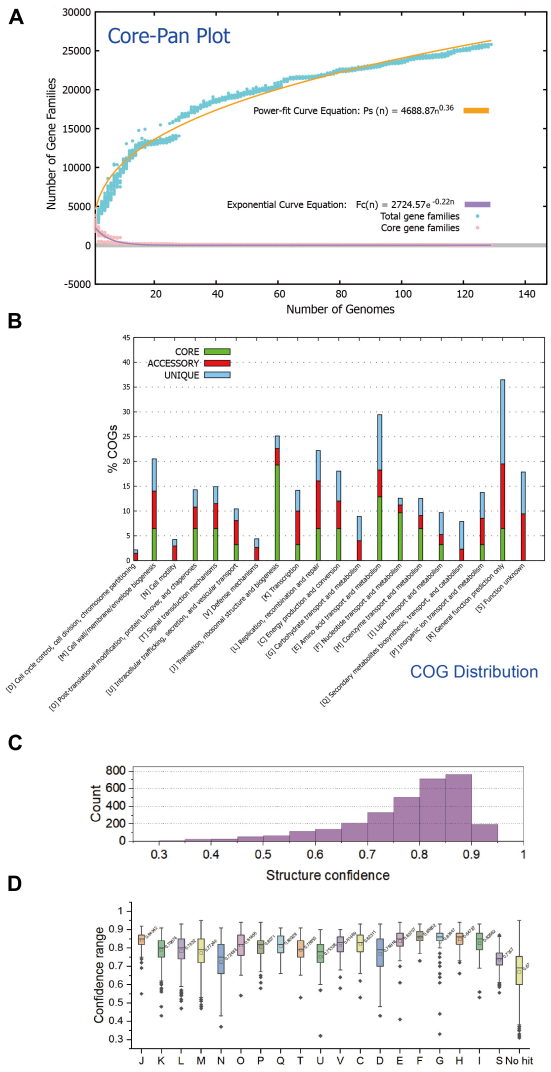


FIGURE 1
The statistical analysis of protein traits and structure confidences of the *Acidithiobacillia* proteome. **(A)** Mathematical modeling of the pangenome and core genome of *Acidithiobacillia*. **(B)** Bar chart showing functional proportions (based on COG categories) of different parts of the *Acidithiobacillia* pangenome (i.e., core, accessory, and unique). **(C)** Histogram showing the distribution of structure confidences of *Acidithiobacillia* pan-proteome. **(D)** Box plot showing the distribution of structure confidences of *Acidithiobacillia* pan-proteome among different COG categories with the average confidence values indicated.

2.2. Highlight of predicted structures

Next, we present and discuss several case-study predictions that focus on unresolved sulfur and iron transport and utilization proteins in *Acidithiobacillia*. These predictions may offer novel insights into the molecular mechanisms of this organism related to biohydrometallurgy. In the Methods section, we provide a summary of the detailed methods employed for these analyses, including substrate binding and molecular dynamics (MD) simulations. It is important to note that the predictions presented here are mainly *de novo*, meaning that no template with more than 30% query identity or covering over 35% of the sequence was available. These predictions can help bridge the knowledge gaps in our understanding of the functional roles and molecular details of these sulfur/iron utilization proteins within the broader biometallurgy system (see Figure 2, with the case-study proteins highlighted in orange rectangles). Although our results have a

10.3389/fmicb.2023.1243987

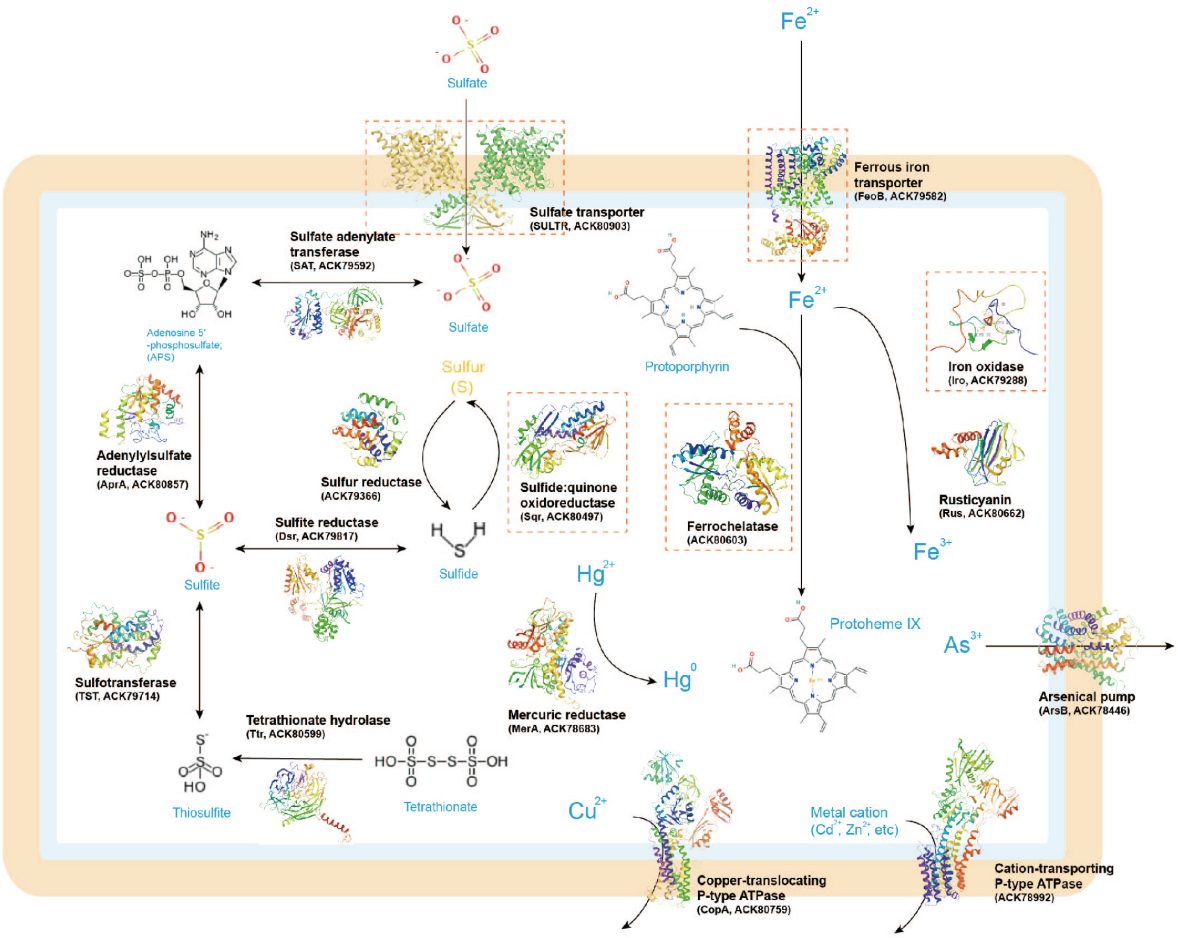


FIGURE 2
A metabolic model of sulfur/iron utilization/biometallurgy related proteins in *Acidithiobacillia* based on genomic inference. The 3D protein structures illustrated in the figure were predicted by AlphaFold2. The case-study proteins discussed in this study are marked with orange rectangles.

significant impact on improving protein structure and function predictions, as well as developing state-of-the-art techniques for high-throughput analysis of large proteomic data, experimental confirmation is ultimately necessary to determine the actual functions of the structure models and the hypothetical key residues within them.

2.2.1. Sulfate transporter

Sulfate (SO_4^{2-}), generated by the microbial oxidation of sulfide minerals, is the dominant aqueous sulfur species in the biomining drainage system and it is vital for microbes to maintain cellular sulfate homeostasis (Borilova et al., 2018). The sulfate transporter is identified in 81.5% of tested *Acidithiobacillia* genomes and has genomic location highly conserved among *Acidithiobacillia* (**Supplementary Figure 2**). The representative protein of it under Genbank (Benson et al., 2018) accession ACK80903 shows 25.32% sequence identity (the best hit in PDB database, HHsearch *p*-value 1.90E−74) to the recently reported chloroplastic sulfate transporter of *Arabidopsis thaliana* (AtSULTR, PDB: 7LHV) (Wang et al., 2021). To our knowledge, no experimental prokaryotic sulfate transporter analog structure currently exists. To alleviate this situation, we obtained the high-confidence modeled structure of ACK80903 (AfSULTR) for comparative structure analysis. The AfSULTR structure exhibits a topology similar to AtSULTR: each monomer is comprised of at least 10 transmembrane (TM) helices followed by a C-terminal anti-Sigma factor antagonist (STAS) domain (**Figure 3A** and **Supplementary Figure 3**). Structure mapping shows that AfSULTR and the AtSULTR monomer have average RMSD of 3.299 angstrom (Å) (**Supplementary Figure 4**). Another important indication of the association between ACK80903 and membrane-anchored transporter is provided by the findings of gene co-occurrence. The analysis revealed that ACK80903 consistently co-occurred with proteins such as proteolipid membrane modulator Pmp3, Na^+/H^+ antiport NhaA, cation transport ATPase (P-type), AI-2E family transporter and FeoC like transcriptional regulator across the comprehensive set of *Acidithiobacillia* genomes (**Supplementary Figure 4**). We built the complete sulfate transporter dimer with two monomers linked by the STAS domains (anchoring on the cytosolic side of the membrane) that swap between the monomers using the dimer structure of AtSULTR (PDB 7LHV) as the dimer template (Wang et al., 2021; **Figure 3B**). A positively charged plane region is visible on the bottom of the TM helices, which is suggested to form electrostatic attachment to the negatively charged microbial membrane (Zhang et al., 2018b; **Figure 3B**). Among the TM helices of AfSULTR, the TM1–7 and TM8–14 (in reference to AtSULTR)

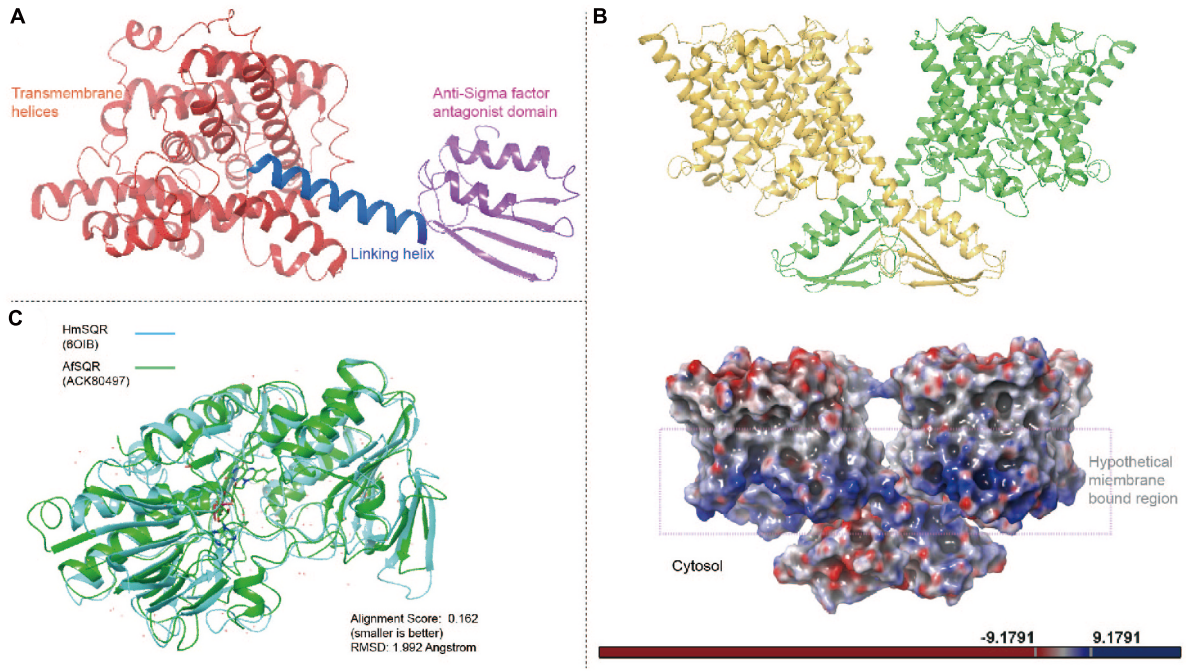


FIGURE 3
The structures of *Acidithiobacillia* sulfate transporter (AfSULTR-ACK80903) and sulfide:quinone oxidoreductase (AfSQR-ACK80497). **(A)** The overall structure of AfSULTR monomer is comprised of transmembrane (TM) helices and anti-Sigma factor antagonist (STAS) domain (shown with red and purple colors, respectively). **(B)** Top panel: the AfSULTR dimer model formed by two identical monomers (shown with yellow and green colors, respectively). Bottom panel: electrostatic potential surfaces of the overall AfSULTR dimer calculated with adaptive Poisson-Boltzmann solver (APBS). **(C)** Superposition of AfSQR-ACK80497 onto human SQR (HmSQR, PDB: 6OIB).

are in a pseudo twofold symmetry, and the TM3 and TM10 arranged in a line are half helices. The crossover region between the N-termini of TM3 and TM10 leaves a crevice (the substrate-binding pocket) surrounded by residues of TM1, TM3, TM8, and TM10 at roughly the center of the TM region (**Supplementary Figure 5a**). These features are consistent with other secondary solute transport proteins (Lu et al., 2011; Alguel et al., 2016; Wang et al., 2019a). The STAS domain of our AfSULTR model is comprised of two α-helices and four β-strands, while the reported AtSULTR counterpart contains four α-helices and four β-strands (Wang et al., 2021). The helix dipoles of TM3 and TM10 carrying the positive electrostatic potential ends seen to orient and attract the negative electrostatic potential of oxygen anions of bound SO_4^{2-} (**Supplementary Figure 5b**). Also, a conserved Arg324 (Arg393 of AtSULTR) from TM10 with positive electrostatic potential was identified to form a putative salt bridge with the bound SO_4^{2-} (**Supplementary Figure 6a**). However, other surrounding residues previously shown to interact with the SO_4^{2-} in the binding pocket in AtSULTR (e.g., Ala153, Phe391, Ser392, Tyr116, and Ser392) (Wang et al., 2021) are all missing in AfSULTR, probably due to protein family diversification. In AfSULTR, the identified surrounding residues (within 6 Å) include three leucine residues (Leu30, Leu369, and Leu379), three valine residues (Val34, Val321, and Val323), two proline residues (Pro69 and Pro71), the above-mentioned Arg324, and Thr70 (**Supplementary Figure 5a**). Leucine and valine contain hydrophobic side chains, which may facilitate the transport of the hydrophilic SO_4^{2-} anion (Abdelraheem et al., 2018). Additionally, a Glu276 in AfSULTR has also been identified at approximately the same position of AtSULTR Glu347 (**Supplementary Figure 6b**), protonation and deprotonation of this residue is suggested to be significant for anion transport and H^+ gradient sensing (Wang et al., 2021).

2.2.2. Sulfide:quinone oxidoreductase

Sulfide:quinone oxidoreductase is a peripheral membrane protein that catalyzes the oxidation of sulfide species to elemental sulfur, belonging to the flavin disulfide reductase (DSR) superfamily (Argyrou and Blanchard, 2004). We find that the protein ortholog of *Acidithiobacillia* under accession ACK80497 putatively represents a novel unresolved SQR-like enzyme, which has a top 1 hit score to human SQR (PDB: 6OIB) (Landry et al., 2019) with 22.51% identity (86% query coverage, *E* value 1e−23). In comparison, ACK80497 shares only 47% query coverage (22.63% identity, *E* value 0.002) with the structure-available SQR homolog of *A. ferrooxidans* (PDB: 3T2Y). The prediction of ACK80497 shows the tandem Rossmann fold repeats commonly seen in the DSR superfamily topology (Argyrou and Blanchard, 2004), with RMSD of 1.992 and 2.207 Å to 6OIB and 3T2Y, respectively (**Figure 3C** and **Supplementary Figure 7**). We compared and identified the conserved triad of active residues in AfSQR-ACK80497, Cys127, Cys158, and Cys331 (equivalent to Cys128, Cys160, and Cys356 of AfSQR-3T2Y, and Cys201, Cys379 of human SQR-6OIB) (**Supplementary Figure 8**). Mutation of these residues was reported to lead to 70∼100% loss of activity (Griesbeck et al., 2002; Cherney et al., 2012). The active site of SQR includes a flavin adenine dinucleotide (FAD) cofactor that accepts and transfers electrons from sulfide species to ubiquinone. However, we failed to identify equivalent residues

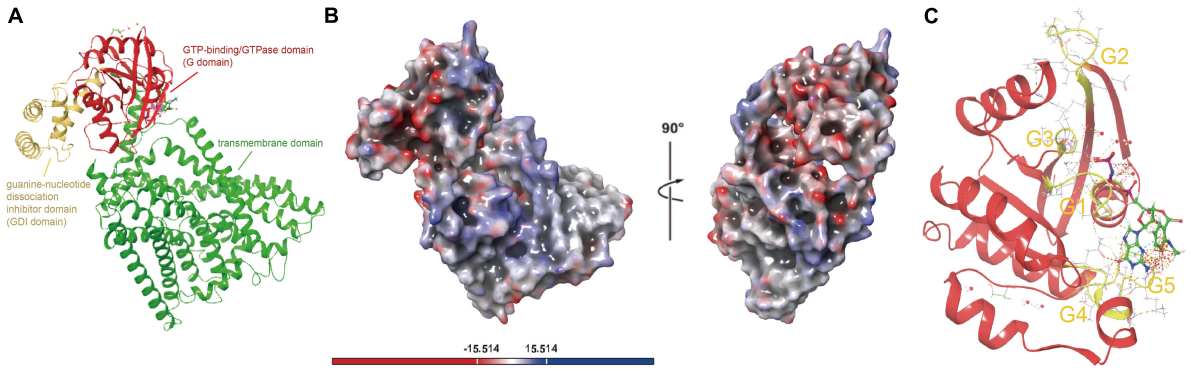


FIGURE 4
Structure visualizations of *Acidithiobacillia* ferrous iron transporter (AfFeoB), iron oxidase and ferrochelatase. **(A)** The overall structure of AfFeoB monomer that consists of G domain, GDI domain, and transmembrane domain. **(B)** Electrostatic potential surfaces of the overall AfFeoB monomer calculated with adaptive Poisson-Boltzmann solver (APBS), which is rotated 90° rightward as indicated to reveal the negatively charged enriched region of AfFeoB that putatively binds the ferric cation. **(C)** The archetypical GTPase motifs G1–G5 (shown with yellow color) in the GTPase domain of AfFeoB that flank the nucleotide-binding pocket.

that were previously shown to bond with the cofactor FAD [e.g., Thr11, Gly12, Ser34, Ala78, Ile302, Gly322, Phe357, and Lys391 in AfSQR-3T2Y (Cherney et al., 2012)], suggesting the existence of a novel ligand-protein interaction diagram in the AfSQR-ACK80497 model. To analyze the putative ligand-protein contact, a classical MD simulation for 15 ns of AfSQR-ACK80497 was performed (**Supplementary Figure 9**), which showed that Met10, Ala39, His43, Gln46, Val81, Lys157, Gly298 of AfSQR-ACK80497 form strong (>50.0% occurrence) direct hydrogen bonds with atoms (e.g., O and N) of the cofactor FAD, while Val44, Ser107, Glu164 of AfSQR-ACK80497 contacts with FAD via water bridges (>50.0% occurrence) (**Supplementary Figure 9**). However, the actual functions of these residues still require further experimental confirmation. Another important indication of the association between ACK80497 and sulfur metabolism is provided by the findings of gene co-occurrence. The analysis revealed that ACK80497 consistently co-occurred with two specific proteins, namely the sulfur carrier protein TusA and the sulfur reduction DsrE/DsrF/DsrH family protein, across the comprehensive set of *Acidithiobacillia* genomes (**Supplementary Figure 9**).

2.2.3. Ferrous iron transporter (FeoB)

Although ferrous iron is one of the primary energy substrates for *Acidithiobacillia,* we still have little knowledge about the uptake process of this substrate. The protein represented by ACK79582 (Genbank accession) encodes a membrane protein FeoB responsible for ferrous iron transport in 94.6% of tested *Acidithiobacillia* (AfFeoB) whose genomic location is highly conserved among *Acidithiobacillia* (**Supplementary Figure 10**). ACK79582 shows only 33% coverage and 35.61% identity (HHsearch *p*-value 3.1E−30) to the PDB hit with top 1 score (PDB 3LX5, NFeoB from *Streptococcus thermophilus*). ACK79582 was then used for structure modeling. The predicted structure of *Acidithiobacillia* (AfFeoB, represented by ACK79582) contains a N-terminal GTP-binding/GTPase domain (G domain, residues 1–169) that shows the canonical G protein fold (a six-stranded β-sheet surrounded by six α-helices), followed by the guanine-nucleotide dissociation inhibitor (GDI) domain (residues 170–257) that consists of a four-helix bundle, which links the GTPase domain and the transmembrane domain (residues 258–766) (**Figure 4A**). A negatively charged enriched region was identified on the surface of AfFeoB (**Figure 4B**), which is suggested to bind the ferric cation. We identified in the GTPase domain of AfFeoB the archetypical GTPase motifs G1–G5 that flank the nucleotide-binding pocket (**Figure 4C** and **Supplementary Figure 11**). These motifs are significant for nucleotide (GTP/GDP) recognition, orientation and reaction catalysis (Scheerer et al., 2008; Guilfoyle et al., 2009). We next compared the residues surrounding the nucleotide-binding site of AfFeoB and other reported FeoB structures (**Supplementary Figure 12**, key conserved residues highlighted with blue rectangles), which identified in AfFeoB the conserved residues Pro10 and Pro56 as essential for maintaining main-chain conformation and affinity for GTP and GDP (Eng et al., 2008), Asp54 and Gly57 that hydrogen-bond with the oxygen of the nucleotide γ-phosphate, Asn11 and Asn115 that contact with the GDP molecule, Asp118 associated with specificity toward the guanine base (Eng et al., 2008), as well as Ala145 and Ser150 in the G5 motif that modulate affinity and release rate of GDP (Guilfoyle et al., 2014a,b). Consistently, the structure of the transmembrane domain of AfFeoB was identified as an analog of the reported concentrative nucleoside transporter vcCNT (PDB 3TIJ) with TM-score 0.59 and RMSD 4.215 Å (**Supplementary Figure 13**; Johnson et al., 2012). Like vcCNT, the transmembrane domain of AfFeoB displayed an overall twofold pseudo-symmetry topology and contained two conserved nucleotide-binding residues, Glu580 and Ser682 (refer to Glu332 and Ser371 at positions similar to that of vcCNT) (Johnson et al., 2012).

2.2.4. Iron oxidase

Iron oxidase (Iro, represented by Genbank ACK79288) in *Acidithiobacillia* is a key protein of the iron respiratory chain that oxidizes ferrous iron, and is closely linked with the biohydrometallurgy efficiency (Zeng et al., 2008). However, the experimental crystal structure of the Iro protein is still lacking. Residues 1–48 of ACK79288 were predicted by SignalP (Almagro Armenteros et al., 2019) to be a TAT(Tat/SPI) type signal

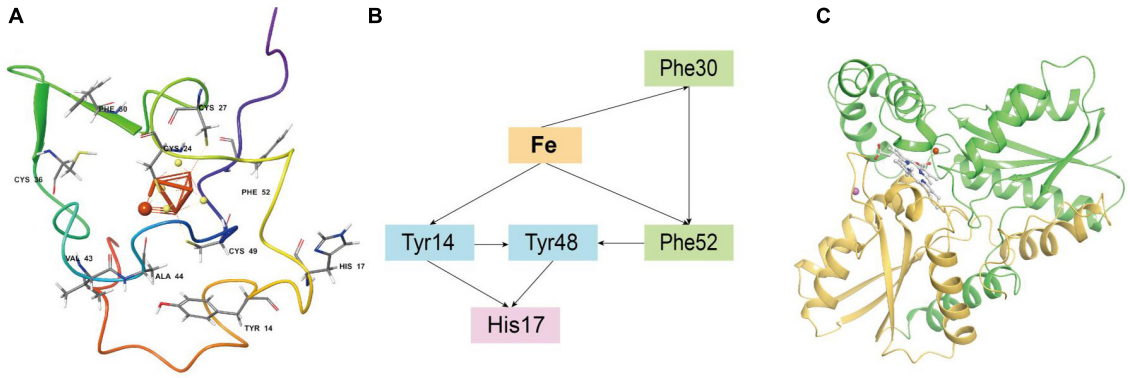


FIGURE 5
(A) The overall structure of *Acidithiobacillia* iron oxidase (Iro) with key conserved cysteine residues and aromatic residues that ligate/stabilize the [Fe(4)S(4)] cluster are shown in ball and stick format, and the free ferrous iron ion is represented by a red sphere. **(B)** Putative electron transfer pathway(s) in Iro identified by Emap using the structure from the last frame of MD simulation. **(C)** The overall topology of *Acidithiobacillia* ferrochelatase (ACK80603) monomer that is comprised of two similar domains (shown with yellow and green colors, respectively). The substrate protoporphyrin molecule is shown in ball and stick format. A free ferrous iron ion and a free magnesium ion are represented by red and pink spheres, respectively.

peptide (**Supplementary Figure 14**) and, therefore, were removed before structure modeling. After modeling, we obtained a high-confidence overall structure of the Iro protein, a side of which was found to be mainly positively charged, forming a putative microbial membrane-bound region (**Supplementary Figure 15**). The [Fe(4)S(4)] cluster is in ligation with four cysteine residues (Cys24, Cys27, Cys36, and Cys49), located in the center of the protein (**Figure 5A**), similar to other HiPIP family proteins (Nogi et al., 2000; Ohno et al., 2017; Kawakami et al., 2021). The [Fe(4)S(4)] cluster is surrounded by the aromatic residues Tyr14, Phe30, and Phe52 (**Figure 5A**), which have been experimentally proven to stabilize the [Fe(4)S(4)] cluster in acid environments (Agarwal et al., 1995; Zeng et al., 2010). Tyr14 especially forms a hydrophobic barrier against solvent attack and mediates electron transfer, substitutions of which may result in protein malfunction (Iwagami et al., 1995). Classical MD simulation (15 ns) reveals that the free ferrous iron to be oxidized is captured and stabilized by Iro mainly through metal coordination effect of atoms from the [Fe(4)S(4)] cluster and ionic interactions from residues Cys24, Val43, and Ala44 (**Supplementary Figure 16**). We further applied Emap (Tazhigulov et al., 2019) to identify putative electron transfer pathway(s) in Iro (**Figure 5B**), which indicated probable electron hopping pathways from ferrous iron to phenylalanine (Phe30 and Phe52), tyrosine (Tyr14 and Tyr48) and histidine (His17) residues.

2.2.5. Ferrochelatase

Ferrochelatase of *Acidithiobacillia* (presented in 98.4% of tested genomes, represented by Genbank ACK80603), involved in cofactor heme biosynthesis, is a membrane-bound protein that catalyzes the insertion of ferrous iron into protoporphyrin IX to form protoheme IX (heme). Its eukaryotic analog is encoded by the nuclear DNA and expressed in the cytoplasm, followed by translocation to the inner mitochondrial membrane, with the active site turned to the mitochondrial matrix (Sellers et al., 2001). In our study, we found that ACK80603 has top 1 hit score to human mitochondrial ferrochelatase (PDB: 2PO7, identity 27.19%, HHsearch *p*-value 1E−62). After modeling, we obtained a high-confidence structure of ACK80603 with an RMSD of 2.182 Å to human mitochondrial ferrochelatase (PDB: 2PO7), which exhibits typical ferrochelatase topology (Sellers et al., 2001), namely a monomer with two similar domains (**Figure 5C**, shown in green and yellow colors, respectively), each containing a Rossmann fold with a four-stranded parallel β-sheet surrounded by α-helices. The two domains are connected by a loop from residues 218–227. We predict a positively charged region in the protein surface that putatively interacts with lipid membranes (**Supplementary Figure 17a**), while negatively charged residues are enriched in the protoporphyrin binding pocket face (**Supplementary Figure 17b**). A previous study generated a reaction model for ferrochelatases based on the data of human ferrochelatase (Sellers et al., 2001), and we attempted to match the critical residues in our structure (**Supplementary Figure 18**):

(1) The conserved carboxylate residues Asp273, Glu276, and Glu280 corresponding to Asp340, Glu343, and Glu347 of human mitochondrial ferrochelatase (PDB 2PO7) putatively form a conduit connecting the active site pocket to the enzyme exterior and participate in proton abstraction from porphyrin (**Supplementary Figure 18**). Replacements of these residues are experimentally proven to hinder the proton abstraction, resulting in no product (heme) formation though the enzyme still binds with protoporphyrin (Sellers et al., 2001);
(2) Ferrous iron is transported from the exterior of the protein via residues Trp163 and Tyr129 (equivalent to Trp227 and Tyr191 of PDB 2PO7) to the site of metalation at residues Arg102 and Tyr103 (equivalent to Arg164 and Tyr165 of PDB 2PO7) centrally located in the active site pocket (**Supplementary Figure 18**), whose role for metalation have been confirmed by mutant tests (Sellers et al., 2001);
(3) The central catalysis residue His195 (refers to His263 of PDB 2PO7) on the opposite side of Arg102 and Tyr103 acts as the proton-acceptor of porphyrin that initializes

metalation in conjunction with proton abstraction. All mutants at His263 (PDB 2PO7) have no measurable enzyme activity (Sellers et al., 2001). Additionally, Trp243 (Trp310 in PDB 2PO7) is involved in saddling of the porphyrin during catalysis (Shi et al., 2006).

Regarding comparisons with microbial ferrochelatase, we also found equivalent residues to Arg115, Tyr123, and Ser130 of *Saccharomyces cerevisiae* ferrochelatase (PDB 2HRE), namely Arg53, Tyr61, and Ser70 in our structure, which putatively form an interaction network with the protoporphyrin substrate as observed previously (Stroupe et al., 2003). Finally, His195 and Glu276 in our structure are located at a position similar to the highly conserved residues His183 and Glu264 in *Bacillus subtilis* ferrochelatase (PDB 2Q2N) that putatively facilitate the insertion reaction of the metal ion into protoporphyrin IX (Hansson et al., 2007). While the conserved active site residues Pro268 and Trp243 (refer to Pro255 and Trp230 of PDB 2Q2N), located in a loop, putatively modulate the regio-specificity of porphyrin binding (Karlberg et al., 2008). MD-based protein-ligand interaction analysis further illustrates that hydrogen-bonding from Arg53, Ser57, Tyr61, Trp65, Ser70, Val275, Glu276, and hydrophobic contact/Pi-Pi stacking from Phe27 take part in stabilization of porphyrin molecule (**Supplementary Figure 19**). Besides, metadynamics analysis shows that ferrous iron stays at the protein conduit without leaving the protein throughout the simulation (20 ns, with ferrous iron firstly placed at the protein entrance position). Free-energy profile of the two collective variables (CVs) depicted in **Supplementary Figure 20** that measure the distances of ferrous iron to metalation residue Tyr103 (C1) and protoporphyrin (N1) throughout simulation (20 ns) displays a wide and deep basin, which indicates that the ferrous iron can be stably captured by the protein (**Supplementary Figure 21**). However, the actual roles of the above-mentioned residues require confirmation by further experimental studies.

2.2.6. Functional predictions of unannotated proteins

A full 35.8% of proteins in the proteome of *Acidithiobacillia* are still labeled as "hypothetical protein" or "domain of unknown function." This microbial "dark matter" awaits exploration, and may also be of significant relevance, especially to the biohydrometallurgy capability of *Acidithiobacillia*. Yet, it is still difficult to crack the mysteries of their functional identity by traditional methods (e.g., genetic manipulation) due to the slow growth rate of this autotrophic organism (Marchand and Silverstein, 2002). Considering that the function of a protein is ultimately defined by its structure, fortunately, many available state-of-the-art deep-learning algorithms can be utilized, such as DeepFRI, who's ability for reliable structure-based function classification and prediction of unknown proteins have been validated (Gligorijević et al., 2021). Thus, we first used the predicted 3D structures of unannotated proteins in the proteome of *Acidithiobacillia* as inputs for DeepFRI (Gligorijević et al., 2021) to perform function prediction. Results show that 93.6% (1,055/1,127) of the unannotated proteins in *Acidithiobacillia* could be assigned structure-based gene ontology (GO) term predictions of cellular components (**Supplementary Table 2**), in which the GO terms cytoplasm (GO:0005737, 29.0%, 306/1,055) and membrane (GO:0016020, 25.8%, 272/1,055) accounted for the largest proportions. A total of 91.3% (1,029/1,127) of unannotated proteins in *Acidithiobacillia* could be assigned structure-based GO term predictions of molecular function (**Supplementary Table 2**), in which the GO terms cellular metabolic process (GO:0044237, 20.0%, 206/1,029) and heterocyclic compound binding (GO:1901363, 10.3%, 106/1,029) accounted for the largest proportions. About 91.8% of these hypothetical proteins have confident scores above the DeepFRI significance cut-off score of 0.5 (**Supplementary Table 2**; Gligorijević et al., 2021), indicating that the predictions are reliable.

Obtaining the general GO-term prediction is only the initial step toward the final characterization of targeted unknown proteins. Other recently published advanced algorithms, such as CHARMM-GUI LBS Finder and Refiner (Guterres et al., 2021) that performs local structure alignment and virtual screening, provide additional tools to identify the putative substrate(s) for an unknown protein using its structural information. For instance, ACK77828 (conserved hypothetical protein) was predicted to be involved in cellular nitrogen compound metabolic process (GO:0034641, score 0.99) by DeepFRI (Gligorijević et al., 2021; **Supplementary Table 2**). Consistent with this, LBS Finder and Refiner (Guterres et al., 2021) predicted the most probable substrate of ACK77828 is a nitrogen-containing compound, namely (2R)-2-amino-3-hydroxysulfanyl-propanoic acid ($C_3H_7NO_3S$, CSO), and CSO was predicted by LBS Finder and Refiner (Guterres et al., 2021) to be bound by ACK77828 in a similar pattern with the transcriptional regulator SarZ (PDB: 3HRM) (**Supplementary Figure 22**).

Robust structure comparison, fold recognition, catalytic site configuration and evolutionary analysis of residues can also be useful during such functional inference. In another case, ACK80741 (conserved hypothetical protein) was given the GO function prediction, disulfide oxidoreductase activity (GO:0015036, score 0.70) by DeepFRI (Gligorijević et al., 2021). Consistent with this, structure comparisons of ACK80741 model (confidence 0.92) with crystalized disulfide oxidoreductase (DSR) family proteins including glutaredoxin (Grx), thioredoxin (Trx), and NrdH show that ACK80741 possesses the combined features of reported DSR proteins (**Figure 6**). ACK80741 exhibits the typical Grx/Trx fold, consisting of a core of four (anti)parallel β-strands flanked by α-helices. In ACK80741, we observed that Lys19 (α-helix-1) forms a hydrogen bond and salt bridge with Thr8 (β-strand-1) and Glu31 (β-strand-2), respectively, and Tyr21 (α-helix-1) and Phe66 (α-helix-3) are involved in an aromatic-aromatic interaction. These cross-helix/strand interactions may be significant for overall structure stability (Lanzarotti et al., 2011) and some of them seem to be unique to ACK80741. Another important indication of the association between ACK80741 and redox hemostasis is provided by the findings of gene co-occurrence. The analysis revealed that ACK80741 consistently co-occurred with respiratory proteins (i.e., respiratory chain assembly protein Aim24 and cytochrome B561), sulfur oxidation *sox* operon protein DUF302 and the stress response protein, Copper binding periplasmic protein CusF, across the comprehensive set of *Acidithiobacillia* genomes (**Supplementary Figure 23**). ACK80741 seems to be closer to the NrdH clade in the phylogenetic tree (**Supplementary Figure 23**). Like NrdH, ACK80741 does not possess the additional N-terminal β-strand/α-helix present in Grx and Trx structures,

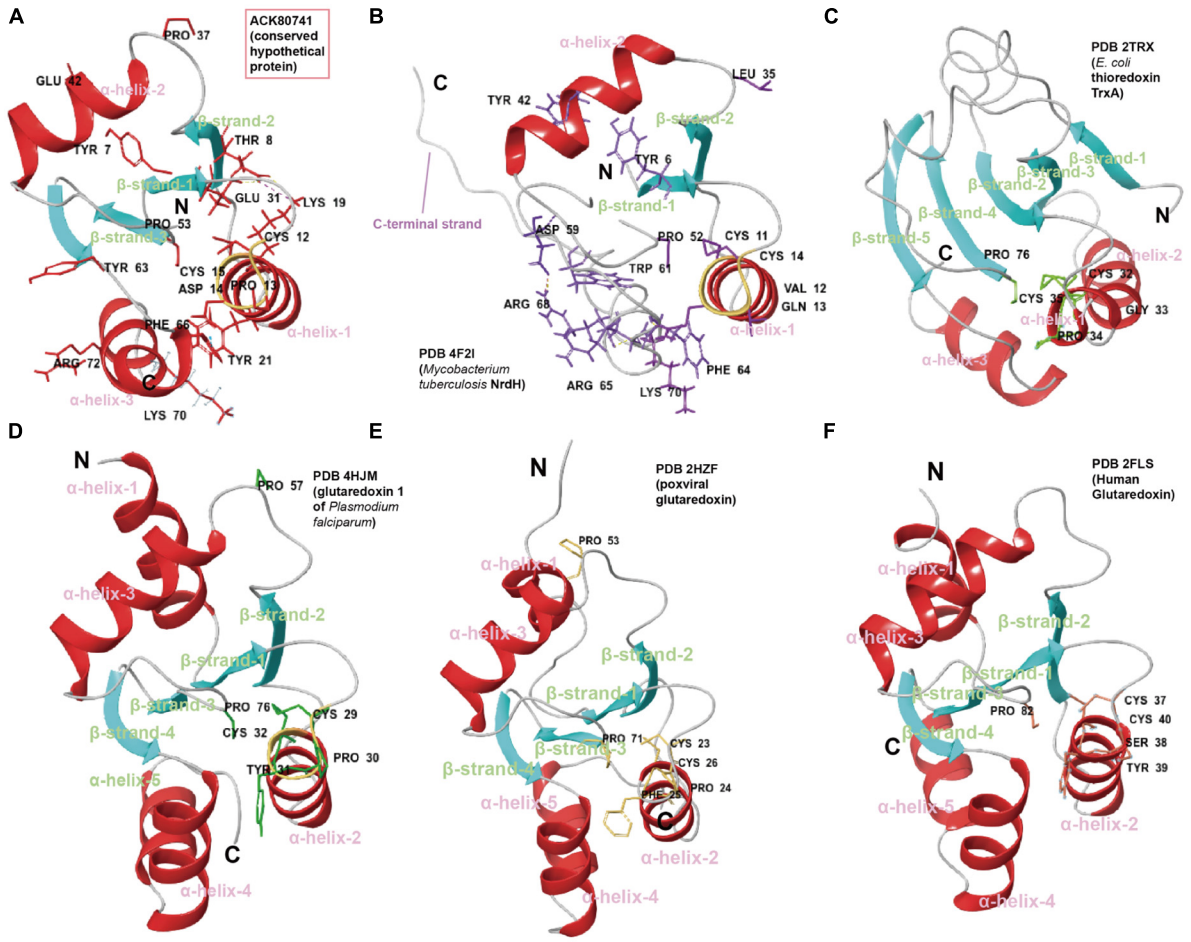


FIGURE 6
Cartoon representations and suggested key residues of **(A)** ACK80741, **(B)** *Mycobacterium tuberculosis* NrdH (PDB 4F2I), **(C)** thioredoxin from *Escherichia coli* TrxA (PDB 2TRX), **(D)** glutaredoxin 1 of *Plasmodium falciparum* (PDB 4HJM), **(E)** poxviral glutaredoxin (PDB 2HZF), and **(F)** human glutaredoxin (PDB 2FLS). The coloring is based on secondary structure (helix, red; sheet, blue; and loop, gray).

while ACK80741 also lack the long C-terminal strand present in NrdH, and the loop regions of ACK80741 are generally shorter than those of Trx. Furthermore, the short α-helices 4 and 5 in Grx structures seem to have merged into a relatively long α-helix in ACK80741 (α-helix-3) (**Figure 6**). We recognize the active site cysteine pair motif (C12-P-D-C15) located in the loop connecting the first β-strand to the second α-helix in ACK80741, which is different from NrdH (CVQC), TrxA (CGPC), and Grx (CPY(F)C) in the residues between the two cysteines (**Supplementary Figure 23**). These residues may affect the redox potential and pKa value of protein (Chivers et al., 1997). In addition to the common turn-inducing Pro53 at the start of β-strand 3, ACK80741 possesses another cis-proline, Pro37 (corresponding to Pro53 of poxviral glutaredoxin, PDB 2HZF) at the start of the third α-helix, which has been shown to be uniquely conserved in orthopoxvirus Grx orthologs (Bacik and Hazes, 2007). ACK80741 also possesses conserved Arg72 and Tyr63, counterparts of MtNrdH Arg68, Trp61 (PDB 4F2I), which are suggested to form a cation-Pi interaction. Lys70, referring to the kink-causing Lys70 of MtNrdH, and Tyr7 that corresponds to Tyr6 of MtNrdH (**Figure 6**; Phulera and Mande, 2013). Further, classical MD simulation (15 ns) reveals that Arg50, Ala52, Thr65, Asp14, Ser9, and Glu68 are important residues involved in substrate (glutathione, GSH) binding and interaction (**Supplementary Figure 24**). Still, the actual functions of the proteins mentioned above require experimental confirmation.

3. Discussion

In this study, we conducted protein structure predictions using AI for *Acidithiobacillia* bacteria, which is commonly employed in biohydrometallurgy. The predictions covered the entire pan-proteome of *Acidithiobacillia*. These structural insights are highly valuable for future research focused on understanding mechanisms and designing proteins. This study builds upon our laboratory's prior investigations into the unique characteristics and applications of *Acidithiobacillia* bacteria (Li et al., 2019; Yang et al., 2020; Tao et al., 2021).

Although continuous efforts have been made to upgrade experimental approaches for protein structure determination, the speed of discovering known structures still lags behind that of sequencing data (Goodsell et al., 2020). Structures of membrane proteins (accounting for only ~1% PDB entries) are particularly

difficult to resolve since they tend to denature and aggregate during purification once removed from their native membrane environment (Bill et al., 2011). In our study, we found that 23.9% of hypothetical proteins in the proteome of *Acidithiobacillia*, as predicted by DeepFRI (Gligorijević et al., 2021), were to be located in the cellular membrane component (highlighted with blue color in **Supplementary Table 2**). The membrane-related proteins play a significant role in various transportation processes and the electron transfer chain, closely associated with iron/sulfur utilization (Yarzábal et al., 2002; Castelle et al., 2008). The elucidation of their structures may shed light on the molecular mechanisms underlying biohydrometallurgy processes in the extreme acidophile *Acidithiobacillia*. Our results demonstrate that the overall prediction reached an average confidence of 0.76 regarding structure prediction of the membrane proteins from *Acidithiobacillia*. This indicates that even for protein classes with limited examples for training datasets, we are able to confidently predict their structures. Additionally, we have highlighted the molecular details of certain membrane proteins (e.g., sulfate transporter) in our case studies. These achievements contribute to why AI-driven accurate protein prediction was selected by the journal *Science* as the top breakthrough of the year 2021 (Thorp, 2021). In our study, more than half of these predictions were of high quality, which is a significant improvement compared to previous studies with less than 40% accuracy (Zhang and Skolnick, 2004).

Additionally, we found that the prediction confidence was independent of the protein sequence length (data not shown), suggesting the capacity of AlphaFold2 (Senior et al., 2020) and RoseTTAFold (Baek et al., 2021) to maintain prediction accuracy, even during the structural prediction of large proteins. This was likely due to the innovation in combining predictions from multiple discontinuous regions to produce an overall structure (Baek et al., 2021), which has outperformed many other modeling programs like Swiss-Model (Waterhouse et al., 2018). These predicted protein structures, when combined with other structure-based analyses, can provide valuable insights into the molecular mechanisms of target proteins and generate scientific hypotheses, including the identification of uncharacterized reaction sites or novel substrate interaction diagrams, as demonstrated in previous studies (Baek et al., 2021; Humphreys et al., 2021; Tunyasuvunakool et al., 2021) and this study (see section "2.2. Highlight of predicted structures").

Although acquiring protein sequence and structural data has become relatively easy, accurately predicting the function of unannotated proteins remains a challenge. In fact, less than 0.8% of the sequences in the UniProt Consortium (2015) have been experimentally characterized and manually annotated in SwissProt (Boutet et al., 2007). Additionally, about 80% of poorly annotated sequences in current databases do not have analogs with similar functions, and 25% of them have no identifiable analogs with a query identity greater than 30% (Zhang et al., 2017). This makes it difficult to perform annotations using traditional homologous transfer approaches. In our study, 35.8% of the proteome from *Acidithiobacillia* consisted of proteins with unknown functions. We propose that the 3D structure of proteins may offer a possible solution to this problem, as the majority of protein domains tend to adopt unique, ordered, and recognizable 3D fold conformations (Das et al., 2015). We were pleased to find that several advanced structure-based high-throughput annotation algorithms are emerging to tackle this challenge, including DeepFRI (Gligorijević et al., 2021), LBS Finder and Refiner (Guterres et al., 2021), CATH (Das et al., 2015), and COFACTOR (Zhang et al., 2017). These algorithms provide valuable information for detailed structure comparison, fold recognition, catalytic site identification, *in silico* reaction simulation, and experimental verification. Some of these algorithms were highlighted in our results (see section "2.2.6. Functional predictions of unannotated proteins"). They leverage structural information, such as the spatial position of amino acids, dihedral angles, the contact matrix representing spatial distances between amino acids, and sub-structure frequency, which have been shown to outperform previous sequence-based prediction methods (Gligorijević et al., 2021). Additionally, gene correlation networks can predict the functions of previously unknown genes based on the functions of adjacent genes (Ma et al., 2018). Other structure-based strategies for enzyme functional characterization have also been demonstrated in previous studies. For example, Zhao et al. (2013) applied large-scale metabolite docking of available 3D protein structures against the KEGG metabolite library and successfully characterized a series of enzymes of unknown functions. Likewise, Hitchcock et al. (2013) proposed the substrate profiles of uncharacterized enzymes by docking metabolites to modeled structures. Finally, Mokrushina et al. (2020) combined structural information with the quantum mechanics/molecular mechanics (QM/MM) method to uncover the catalytic mechanism of an immunoglobulin with novel functionality, which can guide the artificial evolution of valuable enzymes. These strategies significantly expand the possibilities for characterizing and applying unknown proteins.

4. Materials and methods

We have combined protein structural and phylogenetic analyses in this study (**Figure 7**).

4.1. Obtaining sequence of *Acidithiobacillia*

Protein sequences and protein-encoding gene sequences for the proteome of all available *Acidithiobacillia* isolates (n = 129) were downloaded from Genbank (Benson et al., 2018). All protein sequences within the proteome were first manually checked for the presence of ambiguous residue codes (B, J, O, U, Z, or X). Sequence(s) containing these ambiguous residue codes was discarded. Orthofinder v.1.0 with default parameters was used for protein sequence clustering (Emms and Kelly, 2019). The protein sequences were reannotated with eggnog mapper v.2.0 (Cantalapiedra et al., 2021). The size of the *Acidithiobacillia* pangenome was extrapolated by implementing an power law regression function, $Ps = \kappa n^{\gamma}$, using a built-in program of the BPGA v.1.0 (Chaudhari et al., 2016),[1] in which Ps represents the total number of non-orthologous gene families within its pangenome, n represents the number of tested strains, and both κ and γ are free parameters. An exponent γ of <0 suggests

1 https://sourceforge.net/projects/bpgatool/

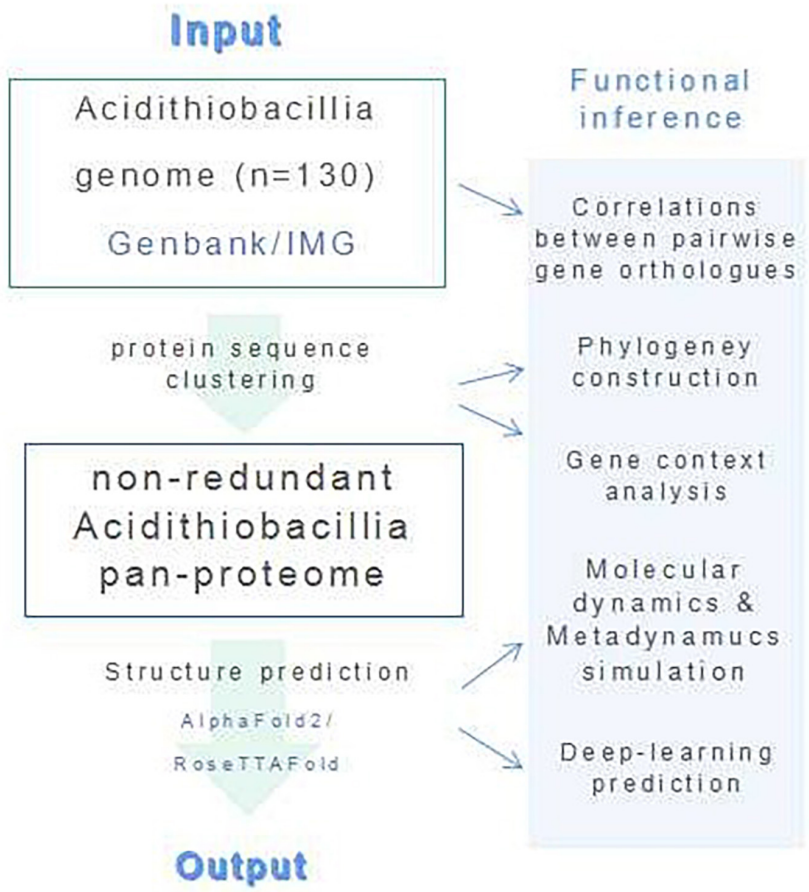


FIGURE 7
A graphical workflow of the computational processes.

the pangenome is "closed," where the size of the pangenome reaches a constant value as extra genomes are added. Conversely, the species is predicted to harbor an "open" pangenome for γ values between 0 and 1. In addition, the size of the core genome was extrapolated by fitting into an exponential decay function, $Fc = \kappa c \exp(-n / \tau c)$, with a built-in program of the BPGA pipeline (Chaudhari et al., 2016), where *Fc* is the number of core gene families, and κc, τc are free parameters. To construct the gene ortholog association network, correlations between pairwise gene orthologs that were present in more than half of the genome were calculated using the CoNet methods in Cytoscape v.3.9.1.[2] Only edges with a significant correlation higher than 0.7 ($p < 0.05$) were retained for network construction. The COG functional categories were assigned by eggNOG-mapper v2 (–evalue 0.001 –score 60 –pident 40 –query_cover 20) after annotation of the query sequences against the COG database (Galperin et al., 2021). We applied Clustal Omega (Sievers and Higgins, 2018) for multiple sequence alignments (MSAs). Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT) (Gerlt, 2017) was used to analyze the gene context in genomes. We used SignalP v.5.0 (Almagro Armenteros et al., 2019)[3] for signal peptide prediction and SOSUI (Hirokawa et al., 1998)[4] for transmembrane region predictions.

4.2. Proteome-scale structure prediction and analysis

We configured the local version of AlphaFold2 (Senior et al., 2020) and RoseTTAFold (Baek et al., 2021) on our laboratory's computation resource, a Dell PowerEdge R940xa server with four Intel Xeon Platinum 8260 processors (total of 148 cores), 1 TB of RAM, installed with Ubuntu 18.04.6 distribution, python 3. Prediction of protein 3D structure was conducted for all checked sequences within the pan-proteome of *Acidithiobacillia* through the local installation. The modeling analysis is generally comprised of six steps: (1) Generate MSAs. (2) Predict secondary structure for HHsearch run. (3) Search for templates. (4) Predict distances and orientations. (5) Perform modeling. (6) Pick final models. The prediction confidences were estimated by multiplying residue-wise accuracy using DeepAccNet (Hiranuma et al., 2021). We applied Visual Molecular Dynamics (VMD) software v.1.9.4

2 https://cytoscape.org

3 http://www.cbs.dtu.dk/services/SignalP/

4 http://harrier.nagahama-i-bio.ac.jp/sosui/

(Humphrey et al., 1996) for structural analysis, visualization, and graphics production. Electrostatic potential was calculated with an adaptive Poisson-Boltzmann solver (APBS) (VMD APBS Plugin, version 1.3.1). TM-align program (Zhang and Skolnick, 2005) was used for structure comparisons. For substrate catalyzing proteins, we applied AutoDock Vina v1.2.1 (Trott and Olson, 2010) to dock the ligand into the predicted structure in reference to its PDB template. All structural predictions generated in this study are available to the community via https://doi.org/10.6084/m9.figshare.19093109.v3.

4.3. Molecular dynamics and metadynamics simulations

Molecular dynamics simulations for the protein-ligand complex were performed using the Desmond Molecular Dynamics System, version 3.6 (D. E. Shaw Research, New York, NY, 2008), with OPLS_2005 force field. We built the simulation system with periodic boundary conditions (PBC), which placed all molecules of the protein-ligand complex in an orthorhombic periodic boundary box with water solvent molecules, together with sodium or chloride ions to balance the systems. Before the production phase, we performed equilibration and energy minimization with the default workflow of Desmond. We conducted MD simulations in an NPT ensemble at a temperature of 300 K and an atmospheric pressure of 1.01325 bar. We integrated the equations of motion with the RESPA integrator, which applied an inner time step of 2.0 fs for bonded and non-bonded interactions within the short-range cut-off and an outer time step of 6.0 fs was used for non-bonded interactions beyond the cut-off. We calculated long-range electrostatic interactions with the Particle-mesh Ewald (PME) method applying a grid spacing of 0.8 A. Additionally, bonds to hydrogen atoms were constrained with the M-SHAKE method. After energy minimization, all molecules were subjected to the final production run for 15–20 ns. The last frame of MD simulation was used as input model for the following metadynamics simulation process using the Desmond Molecular Dynamics System, version 3.6 (D. E. Shaw Research, New York, NY, 2008), for a total of 20 ns. For the metadynamics distance CVs, the Gaussian width was set to 0.05 Å. The starting height of the Gaussian potential was set to 0.03 kcal/mol, and the Gaussians were deposited every 0.09 ps. The simulation was conducted at 300 K and 1.01325 bar pressure. RESPA integrator was applied with a time step of 2.0 fs, and short-range cut-off radius was defined at 9 Å. Electron transfer across the targeted protein was calculated using the structure from the last frame of MD simulation with eMap (Tazhigulov et al., 2019) (default parameters), which applied the graph theory to predict electron tunneling through electron transfer active moieties.

4.4. Phylogenetic tree construction

Phylogenetic tree based on protein sequences was built using PhyML (Guindon et al., 2010) with the Maximum Likelihood (ML) method and 1,000 bootstrap replicates, followed by visualization with iTOL (Letunic and Bork, 2021). Sequences were aligned with MUSCLE (Edgar, 2004) and trimmed with Gblocks (Talavera and Castresana, 2007) prior to tree construction.

5. Conclusion

In this study, by utilizing the advanced AI-driven method, we generated for the first time reliable full-chain structure predictions for the pan-proteome of *Acidithiobacillia*, the model strain for biohydrometallurgy. The median of model confidences was 0.80, and proteins assigned to COG F (nucleotide transport and metabolism), COG H (coenzyme transport and metabolism), and COG J (translation, ribosomal structure and biogenesis) had the highest average confidences. For the convenience of further analyses, the predictions are freely available to the community. In addition, several case studies on structures of conserved sulfur and iron utilization proteins (e.g., sulfate transporter and iron oxidase) that illustrate the effect of high-accuracy predictions are also supplemented. Finally, for the 35.8% unannotated proteins in the proteome, we resorted to the deep-learning algorithm DeepFRI for structure-based functional predictions and successfully obtain GO terms for 93.6% of these unknown proteins. These results pave the way for a better understanding of the biological role of *Acidithiobacillia* in biohydrometallurgy applications.

Data availability statement

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

Author contributions

LL and LZ conceived and designed the research. CJ, ZL, DM, FL, QH, and HY analyzed the data. LL wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Ministry of Science and Technology of China (project no. 2018YFE0110200), the National Natural Science Foundation of China (grants nos. 91851206 and 41877345) and the Key Research and Development Program of Hunan Province (grants nos. 2020WK2022 and 2022SK2076), Fundamental Research Funds for the Central Universities of Central South University (no. 2022ZZTS0420), Hunan International Scientific and Technological Cooperation Base of Environmental Microbiome and Application (No. 2018WK4019), and Natural Science Fundation of Hunan Province (No. 2023JJ30658).

Acknowledgments

We are grateful for resources from the High-Performance Computing Center of Central South 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.

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.2023.1243987/full#supplementary-material

References

Abdelraheem, E. M. M., Shaabani, S., and Dömling, A. (2018). Artificial macrocycles. *Synlett* 29, 1136–1151.

Abergel, C., Nitschke, W., Malarte, G., Bruschi, M., Claverie, J. M., and Giudici-Orticoni, M. T. (2003). The structure of *Acidithiobacillus ferrooxidans* c(4)-cytochrome: A model for complex-induced electron transfer tuning. *Structure* 11, 547–555. doi: 10.1016/s0969-2126(03)00072-8

Agarwal, A., Li, D., and Cowan, J. A. (1995). Role of aromatic residues in stabilization of the [Fe4S4] cluster in high-potential iron proteins (HiPIPs): Physical characterization and stability studies of Tyr-19 mutants of *Chromatium vinosum* HiPIP. *Proc. Natl. Acad. Sci. U.S.A.* 92:9440.

Alguel, Y., Amillis, S., Leung, J., Lambrinidis, G., Capaldi, S., Scull, N. J., et al. (2016). Structure of eukaryotic purine/H(+) symporter UapA suggests a role for homodimerization in transport activity. *Nat. Commun.* 7:11336. doi: 10.1038/ncomms11336

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Argyrou, A., and Blanchard, J. S. (2004). Flavoprotein disulfide reductases: Advances in chemistry and function. *Prog. Nucleic Acid Res. Mol. Biol.* 78, 89–142. doi: 10.1016/S0079-6603(04)78003-4

Bacik, J. P., and Hazes, B. (2007). Crystal structures of a poxviral glutaredoxin in the oxidized and reduced states show redox-correlated structural changes. *J. Mol. Biol.* 365, 1545–1558. doi: 10.1016/j.jmb.2006.11.002

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi: 10.1126/science.abj8754

Banderas, A., and Guiliani, N. (2013). Bioinformatic prediction of gene functions regulated by quorum sensing in the bioleaching bacterium *Acidithiobacillus ferrooxidans*. *Int. J. Mol. Sci.* 14, 16901–16916. doi: 10.3390/ijms140816901

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al. (2018). GenBank. *Nucleic Acids Res.* 46, D41–D47.

Bill, R. M., Henderson, P. J. F., Iwata, S., Kunji, E. R. S., Michel, H., Neutze, R., et al. (2011). Overcoming barriers to membrane protein structure determination. *Nat. Biotechnol.* 29, 335–340.

Borilova, S., Mandl, M., Zeman, J., Kucera, J., Pakostova, E., Janiczek, O., et al. (2018). Can sulfate be the first dominant aqueous sulfur species formed in the oxidation of pyrite by *Acidithiobacillus ferrooxidans*? *Front. Microbiol.* 9:3134. doi: 10.3389/fmicb.2018.03134

Botuyan, M. V., Toy-Palmer, A., Chung, J., Blake, R. C. II, Beroza, P., Case, D. A., et al. (1996). NMR solution structure of Cu(I) rusticyanin from *Thiobacillus ferrooxidans*: Structural basis for the extreme acid stability and redox potential. *J. Mol. Biol.* 263, 752–767. doi: 10.1006/jmbi.1996.0613

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods Mol. Biol.* 406, 89–112.

Bruscella, P., Cassagnaud, L., Ratouchniak, J., Brasseur, G., Lojou, E., Amils, R., et al. (2005). The HiPIP from the acidophilic *Acidithiobacillus ferrooxidans* is correctly processed and translocated in *Escherichia coli*, in spite of the periplasm pH difference between these two micro-organisms. *Microbiology* 151, 1421–1431. doi: 10.1099/mic.0.27476-0

Campodonico, M. A., Vaisman, D., Castro, J. F., Razmilic, V., Mercado, F., Andrews, B. A., et al. (2016). *Acidithiobacillus ferrooxidans's* comprehensive model driven analysis of the electron transfer metabolism and synthetic strain design for biomining applications. *Metab. Eng. Commun.* 3, 84–96. doi: 10.1016/j.meteno.2016.03.003

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

Castelle, C., Guiral, M., Malarte, G., Ledgham, F., Leroy, G., Brugna, M., et al. (2008). A new iron-oxidizing/O2-reducing supercomplex spanning both inner and outer membranes, isolated from the extreme acidophile *Acidithiobacillus ferrooxidans*. *J. Biol. Chem.* 283, 25803–25811. doi: 10.1074/jbc.M802496200

Chaudhari, N. M., Gupta, V. K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* 6:24373. doi: 10.1038/srep24373

Cherney, M. M., Zhang, Y., James, M. N., and Weiner, J. H. (2012). Structure-activity characterization of sulfide:quinone oxidoreductase variants. *J. Struct. Biol.* 178, 319–328. doi: 10.1016/j.jsb.2012.04.007

Cherney, M. M., Zhang, Y., Solomonson, M., Weiner, J. H., and James, M. N. (2010). Crystal structure of sulfide:quinone oxidoreductase from *Acidithiobacillus ferrooxidans*: Insights into sulfidotrophic respiration and detoxification. *J. Mol. Biol.* 398, 292–305. doi: 10.1016/j.jmb.2010.03.018

Chivers, P. T., Prehoda, K. E., and Raines, R. T. (1997). The CXXC motif: A rheostat in the active site. *Biochemistry* 36, 4061–4066.

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., and Orengo, C. A. (2015). Functional classification of CATH superfamilies: A domain-based approach for protein function annotation. *Bioinformatics* 31, 3460–3467.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

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

Eng, E. T., Jalilian, A. R., Spasov, K. A., and Unger, V. M. (2008). Characterization of a novel prokaryotic GDP dissociation inhibitor domain from the G protein coupled membrane protein FeoB. *J. Mol. Biol.* 375, 1086–1097. doi: 10.1016/j.jmb.2007.11.027

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., and Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 49, D274–D281. doi: 10.1093/nar/gkaa1018

Gerlt, J. A. (2017). Genomic enzymology: Web tools for leveraging protein family sequence-function space and genome context to discover novel functions. *Biochemistry* 56, 4293–4308. doi: 10.1021/acs.biochem.7b00614

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* 12, 3168–3168.

Goodsell, D. S., Zardecki, C., Di Costanzo, L., Duarte, J. M., Hudson, B. P., Persikova, I., et al. (2020). RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.* 29, 52–65.

Griesbeck, C., Schütz, M., Schödl, T., Bathe, S., Nausch, L., Mederer, N., et al. (2002). Mechanism of sulfide-quinone reductase investigated using site-directed mutagenesis and sulfur analysis. *Biochemistry* 41, 11552–11565. doi: 10.1021/bi026032b

Guan, J., Berlinger, S. A., Li, X., Chao, Z., Sousa, E. S. V., Banta, S., et al. (2017). Development of reactor configurations for an electrofuels platform utilizing genetically modified iron oxidizing bacteria for the reduction of CO_2 to biochemicals. *J. Biotechnol.* 245, 21–27. doi: 10.1016/j.jbiotec.2017.02.004

Guilfoyle, A., Maher, M. J., Rapp, M., Clarke, R., Harrop, S., and Jormakka, M. (2009). Structural basis of GDP release and gating in G protein coupled $Fe2^{+}$ transport. *Embo J.* 28, 2677–2685. doi: 10.1038/emboj.2009.208

Guilfoyle, A. P., Deshpande, C. N., Schenk, G., Maher, M. J., and Jormakka, M. (2014a). Exploring the correlation between the sequence composition of the nucleotide binding G5 loop of the FeoB GTPase domain (NFeoB) and intrinsic rate of GDP release. *Biosci. Rep.* 34:e00158. doi: 10.1042/BSR20140152

Guilfoyle, A. P., Deshpande, C. N., Vincent, K., Pedroso, M. M., Schenk, G., Maher, M. J., et al. (2014b). Structural and functional analysis of a FeoB A143S G5 loop mutant explains the accelerated GDP release rate. *FEBS J.* 281, 2254–2265. doi: 10.1111/febs.12779

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guterres, H., Park, S.-J., Zhang, H., and Im, W. (2021). CHARMM-GUI LBS finder & refiner for ligand binding site prediction and refinement. *J. Chem. Inf. Model.* 61, 3744–3751.

Hansson, M. D., Karlberg, T., Rahardja, M. A., Al-Karadaghi, S., and Hansson, M. (2007). Amino acid residues His183 and Glu264 in *Bacillus subtilis* ferrochelatase direct and facilitate the insertion of metal ion into protoporphyrin IX. *Biochemistry* 46, 87–94. doi: 10.1021/bi061760a

Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., and Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* 12:1340. doi: 10.1038/s41467-021-21511-x

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14, 378–379. doi: 10.1093/bioinformatics/14.4.378

Hitchcock, D. S., Fan, H., Kim, J., Vetting, M., Hillerich, B., Seidel, R. D., et al. (2013). Structure-guided discovery of new deaminase enzymes. *J. Am. Chem. Soc.* 135, 13927–13933. doi: 10.1021/ja4066078

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *J. Mol. Graph* 14, 27–28.

Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., et al. (2021). Computed structures of core eukaryotic protein complexes. *Science* 374, eabm4805.

Inaba, Y., West, A. C., and Banta, S. (2020). Enhanced microbial corrosion of stainless steel by *Acidithiobacillus* ferrooxidans through the manipulation of substrate oxidation and overexpression of rus. *Biotechnol. Bioeng.* 117, 3475–3485. doi: 10.1002/bit.27509

Iwagami, S. G., Creagh, A. L., Haynes, C. A., Borsari, M., Felli, I. C., Piccioli, M., et al. (1995). The role of a conserved tyrosine residue in high-potential iron sulfur proteins. *Protein Sci.* 4, 2562–2572.

Johnson, Z. L., Cheong, C.-G., and Lee, S.-Y. (2012). Crystal structure of a concentrative nucleoside transporter from *Vibrio cholerae* at 2.4 Å. *Nature* 483, 489–493. doi: 10.1038/nature10882

Kanao, T., Hase, N., Nakayama, H., Yoshida, K., Nishiura, K., Kosaka, M., et al. (2021). Reaction mechanism of tetrathionate hydrolysis based on the crystal structure of tetrathionate hydrolase from *Acidithiobacillus ferrooxidans*. *Protein Sci.* 30, 328–338. doi: 10.1002/pro.3984

Karlberg, T., Hansson, M. D., Yengo, R. K., Johansson, R., Thorvaldsen, H. O., Ferreira, G. C., et al. (2008). Porphyrin binding and distortion and substrate specificity in the ferrochelatase reaction: The role of active site residues. *J. Mol. Biol.* 378, 1074–1083. doi: 10.1016/j.jmb.2008.03.040

Kawakami, T., Yu, L. J., Liang, T., Okazaki, K., Madigan, M. T., Kimura, Y., et al. (2021). Crystal structure of a photosynthetic LH1-RC in complex with its electron donor HiPIP. *Nat. Commun.* 12:1104. doi: 10.1038/s41467-021-21397-9

Landry, A. P., Moon, S., Kim, H., Yadav, P. K., Guha, A., Cho, U. S., et al. (2019). A catalytic trisulfide in human sulfide quinone oxidoreductase catalyzes coenzyme a persulfide synthesis and inhibits butyrate oxidation. *Cell Chem. Biol.* 26, 1515–1525.e. doi: 10.1016/j.chembiol.2019.09.010

Lanzarotti, E., Biekofsky, R. R., Estrin, D. A., Marti, M. A., and Turjanski, A. G. (2011). Aromatic-aromatic interactions in proteins: Beyond the dimer. *J. Chem. Inf. Model* 51, 1623–1633. doi: 10.1021/ci200062e

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, L., Liu, Z., Meng, D., Liu, X., Li, X., Zhang, M., et al. (2019). Comparative genomic analysis reveals the distribution, organization, and evolution of metal resistance genes in the genus *Acidithiobacillus*. *Appl. Environ. Microbiol.* 85, e2153–e2118. doi: 10.1128/AEM.02153-18

Lin, Y. (2018). What's happened over the last five years with high-throughput protein crystallization screening? *Expert. Opin. Drug Discov.* 13, 691–695. doi: 10.1080/17460441.2018.1465924

Liu, W., Lin, J., Pang, X., Mi, S., Cui, S., and Lin, J. (2013). Increases of ferrous iron oxidation activity and arsenic stressed cell growth by overexpression of Cyc2 in *Acidithiobacillus ferrooxidans* ATCC19859. *Biotechnol. Appl. Biochem.* 60, 623–628. doi: 10.1002/bab.1110

Lu, F., Li, S., Jiang, Y., Jiang, J., Fan, H., Lu, G., et al. (2011). Structure and mechanism of the uracil transporter UraA. *Nature* 472, 243–246.

Ma, B., Zhao, K., Lv, X., Su, W., Dai, Z., Gilbert, J. A., et al. (2018). Genetic correlation network prediction of forest soil microbial functional organization. *ISME J.* 12, 2492–2505. doi: 10.1038/s41396-018-0232-8

Marchand, E. A., and Silverstein, J. (2002). Influence of heterotrophic microbial growth on biological oxidation of pyrite. *Environ. Sci. Technol.* 36, 5483–5490.

Mokrushina, Y. A., Golovin, A. V., Smirnov, I. V., Chatziefthimiou, S. D., Stepanova, A. V., Bobik, T. V., et al. (2020). Multiscale computation delivers organophosphorus reactivity and stereoselectivity to immunoglobulin scavengers. *Proc. Natl. Acad. Sci. U. S. A.* 117, 22841–22848. doi: 10.1073/pnas.2010317117

Moya-Beltrán, A., Beard, S., Rojas-Villalobos, C., Issotta, F., Gallardo, Y., Ulloa, R., et al. (2021). Genomic evolution of the class *Acidithiobacillia*: Deep-branching *Proteobacteria* living in extreme acidic conditions. *ISME J.* 15, 3221–3238. doi: 10.1038/s41396-021-00995-x

Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T., and Miki, K. (2000). Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from *Thermochromatium tepidum*: Thermostability and electron transfer. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13561–13566. doi: 10.1073/pnas.240224997

Ohno, H., Takeda, K., Niwa, S., Tsujinaka, T., Hanazono, Y., Hirano, Y., et al. (2017). Crystallographic characterization of the high-potential iron-sulfur protein in the oxidized state at 0.8 Å resolution. *PLoS One* 12, e0178183. doi: 10.1371/journal.pone.0178183

Parte, A. C. (2014). LPSN–list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res.* 42, D613–D616. doi: 10.1093/nar/gkt1111

Phulera, S., and Mande, S. C. (2013). The crystal structure of mycobacterium tuberculosis NrdH at 0.87 Å suggests a possible mode of its activity. *Biochemistry* 52, 4056–4065. doi: 10.1021/bi400191z

Quatrini, R., Appia-Ayme, C., Denis, Y., Jedlicki, E., Holmes, D. S., and Bonnefoy, V. (2009). Extending the models for iron and sulfur oxidation in the extreme Acidophile *Acidithiobacillus ferrooxidans*. *BMC Genomics* 10:394. doi: 10.1186/1471-2164-10-394

Ramírez, P., Guiliani, N., Valenzuela, L., Beard, S., and Jerez, C. A. (2004). Differential protein expression during growth of *Acidithiobacillus* ferrooxidans on ferrous iron, sulfur compounds, or metal sulfides. *Appl. Environ. Microbiol.* 70, 4491–4498.

Scheerer, P., Park, J. H., Hildebrand, P. W., Kim, Y. J., Krauss, N., Choe, H. W., et al. (2008). Crystal structure of opsin in its G-protein-interacting conformation. *Nature* 455, 497–502. doi: 10.1038/nature07330

Sellers, V. M., Wu, C. K., Dailey, T. A., and Dailey, H. A. (2001). Human ferrochelatase: Characterization of substrate-iron binding and proton-abstracting residues. *Biochemistry* 40, 9821–9827. doi: 10.1021/bi010012c

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.

Shi, Z., Franco, R., Haddad, R., Shelnutt, J. A., and Ferreira, G. C. (2006). The conserved active-site loop residues of ferrochelatase induce porphyrin conformational changes necessary for catalysis. *Biochemistry* 45, 2904–2912. doi: 10.1021/bi051907i

Sievers, F., and Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27, 135–145.

Stroupe, M. E., Leech, H. K., Daniels, D. S., Warren, M. J., and Getzoff, E. D. (2003). CysG structure reveals tetrapyrrole-binding features and novel regulation of siroheme biosynthesis. *Nat. Struct. Biol.* 10, 1064–1073. doi: 10.1038/nsb1007

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

Tao, J., Liu, X., Luo, X., Teng, T., Jiang, C., Drewniak, L., et al. (2021). An integrated insight into bioleaching performance of chalcopyrite mediated by microbial factors: Functional types and biodiversity. *Bioresour. Technol.* 319:124219. doi: 10.1016/j.biortech.2020.124219

Tazhigulov, R. N., Gayvert, J. R., Wei, M., and Bravaya, K. B. (2019). eMap: A web application for identifying and visualizing electron or hole hopping pathways in proteins. *J. Phys. Chem. B* 123, 6946–6951.

Thorp, H. H. (2021). Proteins, proteins everywhere. *Science* 374:1415.

Trott, O., and Olson, A. J. (2010). AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596.

UniProt Consortium (2015). UniProt: A hub for protein information. *Nucleic Acids Res.* 43, D204–D212.

Valdés, J., Pedroso, I., Quatrini, R., Dodson, R. J., Tettelin, H., Blake, R. II, et al. (2008). *Acidithiobacillus ferrooxidans* metabolism: From genome sequence to industrial applications. *BMC Genomics* 9:597. doi: 10.1186/1471-2164-9-597

Vera, M., Krok, B., Bellenberg, S., Sand, W., and Poetsch, A. (2013). Shotgun proteomics study of early biofilm formation process of *Acidithiobacillus ferrooxidans* ATCC 23270 on pyrite. *Proteomics* 13, 1133–1144. doi: 10.1002/pmic.201200386

Walter, R. L., Ealick, S. E., Friedman, A. M., Blake, R. C. II, Proctor, P., and Shoham, M. (1996). Multiple wavelength anomalous diffraction (MAD) crystal structure of rusticyanin: A highly oxidizing cupredoxin with extreme acid stability. *J. Mol. Biol.* 263, 730–751. doi: 10.1006/jmbi.1996.0612

Wang, R., Lin, J.-Q., Liu, X.-M., Pang, X., Zhang, C.-J., Yang, C.-L., et al. (2019b). Sulfur oxidation in the acidophilic autotrophic *Acidithiobacillus* spp. *Front. Microbiol.* 9, 3290–3290. doi: 10.3389/fmicb.2018.03290

Wang, C., Sun, B., Zhang, X., Huang, X., Zhang, M., Guo, H., et al. (2019a). Structural mechanism of the active bicarbonate transporter from cyanobacteria. *Nat. Plants* 5, 1184–1193.

Wang, J., Bai, J., Xu, J., and Liang, B. (2009). Bioleaching of metals from printed wire boards by *Acidithiobacillus ferrooxidans* and *Acidithiobacillus thiooxidans* and their mixture. *J. Hazard Mater.* 172, 1100–1105. doi: 10.1016/j.jhazmat.2009.07.102

Wang, L., Chen, K., and Zhou, M. (2021). Structure and function of an *Arabidopsis thaliana* sulfate transporter. *Nat. Commun.* 12:4455. doi: 10.1038/s41467-021-24778-2

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303.

Yang, B., Zhao, C., Luo, W., Liao, R., Gan, M., Wang, J., et al. (2020). Catalytic effect of silver on copper release from chalcopyrite mediated by *Acidithiobacillus ferrooxidans*. *J. Hazard Mater.* 392:122290. doi: 10.1016/j.jhazmat.2020.122290

Yarzábal, A., Brasseur, G., Ratouchniak, J., Lund, K., Lemesle-Meunier, D., DeMoss, J. A., et al. (2002). The high-molecular-weight cytochrome c Cyc2 of *Acidithiobacillus ferrooxidans* is an outer membrane protein. *J. Bacteriol.* 184, 313–317. doi: 10.1128/JB.184.1.313-317.2002

Zeng, J., Jiang, H., Liu, Y., Liu, J., and Qiu, G. (2008). Expression, purification and characterization of a high potential iron-sulfur protein from *Acidithiobacillus ferrooxidans*. *Biotechnol. Lett.* 30, 905–910. doi: 10.1007/s10529-007-9612-2

Zeng, J., Liu, Q., Zhang, X., Mo, H., Wang, Y., Chen, Q., et al. (2010). Functional roles of the aromatic residues in the stabilization of the [Fe(4)S(4)] cluster in the Iro protein from *Acidithiobacillus ferrooxidans*. *J. Microbiol. Biotechnol.* 20, 294–300.

Zhang, C., Freddolino, P. L., and Zhang, Y. (2017). COFACTOR: Improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* 45, W291–W299. doi: 10.1093/nar/gkx366

Zhang, S., Yan, L., Xing, W., Chen, P., Zhang, Y., and Wang, W. (2018a). *Acidithiobacillus ferrooxidans* and its potential application. *Extremophiles* 22, 563–579. doi: 10.1007/s00792-018-1024-9

Zhang, X., Geng, H., Gong, L., Zhang, Q., Li, H., Zhang, X., et al. (2018b). Modification of the surface of titanium with multifunctional chimeric peptides to prevent biofilm formation via inhibition of initial colonizers. *Int. J. Nanomed.* 13, 5361–5375. doi: 10.2147/IJN.S170819

Zhang, X., Liu, X., Li, L., Wei, G., Zhang, D., Liang, Y., et al. (2019). Phylogeny, divergent evolution, and speciation of sulfur-oxidizing *Acidithiobacillus* populations. *BMC Genomics* 20:438. doi: 10.1186/s12864-019-5827-6

Zhang, Y., and Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7594–7599.

Zhang, Y., and Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33, 2302–2309. doi: 10.1093/nar/gki524

Zhao, S., Kumar, R., Sakai, A., Vetting, M. W., Wood, B. M., Brown, S., et al. (2013). Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* 502, 698–702.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 12 September 2023
DOI 10.3389/fmicb.2023.1211999

Check for updates

OPEN ACCESS

EDITED BY
Satyanarayana Tulasi,
University of Delhi, India

REVIEWED BY
Prabhu B. Patil,
Institute of Microbial Technology (CSIR), India
Eduardo Abreo,
National Institute for Agricultural Research (INIA), Uruguay

*CORRESPONDENCE
Kirill S. Antonets
✉ k.antonets@arriam.ru

RECEIVED 25 April 2023
ACCEPTED 21 August 2023
PUBLISHED 12 September 2023

CITATION
Shikov AE, Merkushova AV, Savina IA, Nizhnikov AA and Antonets KS (2023) The man, the plant, and the insect: shooting host specificity determinants in *Serratia marcescens* pangenome. *Front. Microbiol.* 14:1211999. doi: 10.3389/fmicb.2023.1211999

COPYRIGHT
© 2023 Shikov, Merkushova, Savina, Nizhnikov and Antonets. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The man, the plant, and the insect: shooting host specificity determinants in *Serratia marcescens* pangenome

Anton E. Shikov[1,2], Anastasiya V. Merkushova[1], Iuliia A. Savina[1], Anton A. Nizhnikov[1,2] and Kirill S. Antonets[1,2]*

[1]Laboratory for Proteomics of Supra-Organismal Systems, All-Russia Research Institute for Agricultural Microbiology, St. Petersburg, Russia, [2]Faculty of Biology, St. Petersburg State University, St. Petersburg, Russia

Introduction: *Serratia marcescens* is most commonly known as an opportunistic pathogen causing nosocomial infections. It, however, was shown to infect a wide range of hosts apart from vertebrates such as insects or plants as well, being either pathogenic or growth-promoting for the latter. Despite being extensively studied in terms of virulence mechanisms during human infections, there has been little evidence of which factors determine *S. marcescens* host specificity. On that account, we analyzed *S. marcescens* pangenome to reveal possible specificity factors.

Methods: We selected 73 high-quality genome assemblies of complete level and reconstructed the respective pangenome and reference phylogeny based on core genes alignment. To find an optimal pipeline, we tested current pangenomic tools and obtained several phylogenetic inferences. The pangenome was rich in its accessory component and was considered open according to the Heaps' law. We then applied the pangenome-wide associating method (pan-GWAS) and predicted positively associated gene clusters attributed to three host groups, namely, humans, insects, and plants.

Results: According to the results, significant factors relating to human infections included transcriptional regulators, lipoproteins, ABC transporters, and membrane proteins. Host preference toward insects, in its turn, was associated with diverse enzymes, such as hydrolases, isochorismatase, and N-acetyltransferase with the latter possibly exerting a neurotoxic effect. Finally, plant infection may be conducted through type VI secretion systems and modulation of plant cell wall synthesis. Interestingly, factors associated with plants also included putative growth-promoting proteins like enzymes performing xenobiotic degradation and releasing ammonium irons. We also identified overrepresented functional annotations within the sets of specificity factors and found that their functional characteristics fell into separate clusters, thus, implying that host adaptation is represented by diverse functional pathways. Finally, we found that mobile genetic elements bore specificity determinants. In particular, prophages were mainly associated with factors related to humans, while genetic islands-with insects and plants, respectively.

Discussion: In summary, functional enrichments coupled with pangenomic inferences allowed us to hypothesize that the respective host preference is carried out through distinct molecular mechanisms of virulence. To the best of our knowledge, the presented research is the first to identify specific genomic

features of *S. marcescens* assemblies isolated from different hosts at the pangenomic level.

KEYWORDS

***Serratia marcescens*, pangenome, host specificity, virulence, pan-GWAS, human infections, PGP**

1. Introduction

A rod-shaped Gram-negative bacterium, *Serratia marcescens*, belonging to the *Enterobacteriaceae* family, is ubiquitously found in various environmental niches, including soil, freshwater, and air (Matteoli et al., 2021) and often isolated from animals and plants as well (Mahlen, 2011). The description of the species dates back to 1,819 in Italy as red spots on polenta (Merlino, 1924). Since this bacterium was initially considered non-pathogenic and produced red pigment, it found wide application as a biological warfare test organism, which led to the release of the bacterium by the US government from the 1940s to the 1960s with a view of monitoring possible bioterrorism threats (Mahlen, 2011). Later on, however, it has become clear that *S. marcescens* is a causative agent of nosocomial infections, especially in neonates and immunocompromised patients (Moradigaravand et al., 2016; Saralegui et al., 2020). Clinical studies illustrated that *S. marcescens* is capable of invading the respiratory (Dessì et al., 2009) and urinary tracts (Lancaster, 1962), wounds (Moradigaravand et al., 2016), and bloodstream (Haddy et al., 1996). It is worth mentioning that bacteremia, in the case of bloodstream invasion, is a severe disease with a mortality rate exceeding 20% (Haddy et al., 1996). The half-century history of epidemiological monitoring brought evidence of multiple infection outbreaks caused by the bacterium (Escribano et al., 2019; Saralegui et al., 2020). Due to the emergence of antibiotic-resistant strains driven by HGT (horizontal gene transfer), studying genomic features of pathogenic *S. marcescens* isolates is of great importance for the healthcare system (Matteoli et al., 2021; Shikov et al., 2022). Several *S. marcescens* strains are pathogenic to bovines (Friman et al., 2019), birds (Saidenberg et al., 2007), reptiles (Pye et al., 1999), and fishes (Baya et al., 1992).

The repertoire of hosts affected by *S. marcescens* is not limited to vertebrates. Being present in the insect gut, it could be efficiently transmitted between individuals (Dupriez et al., 2022). Virulence potential is delineated by host species, development stage, and environmental conditions (Dupriez et al., 2022). The severity of the disease also depends on the abundance of *S. marcescens* in the host's microbiome (Sikorowski et al., 2001). Apart from insecticidal activity, *S. marcescens* exerts an effect on plants which can be either pathogenic or stimulatory. Recent studies have provided multiple evidence that *S. marcescens* is an opportunistic plant pathogen inducing black rot of sweet orange (*Citrus sinensis* L.) (Hasan et al., 2020), cucurbit yellow vine disease (CYVD) (Besler and Little, 2017), necrotic lesions on oleander (*Nerium oleander* L.) leaves (Fodor et al., 2022), bell pepper (*Capsicum annuum* L.) soft-rot (Gillis et al., 2014). On the other hand, some *S. marcescens* strains exhibit plant growth-promoting (PGP) properties (Matteoli et al., 2018; Abreo and Altier, 2019). These features include induction of phosphorus solubilization (Tripura et al., 2007), defense from insects by chitinases (Vaikuntapu et al., 2016) coupled with prodigiosin biocontrol (Suryawanshi et al., 2015), and antagonism with fungal plant pathogens (Li et al., 2015; Matteoli et al., 2018).

Studying the pathogenicity of *S. marcescens* as a model object has expanded our understanding of the infection process mediated by various virulence factors. These included secreted proteases (Kamata et al., 1985), extracellular phospholipases (Shimuta et al., 2009), hemolysins (ShlA) (Abreo and Altier, 2019), iron acquisition transporters (Létoffé et al., 1994), and lipopolysaccharides (Kurz et al., 2003). At the same time, the mechanisms affecting specificity and adaptation to certain hosts are poorly studied. General functional peculiarities of putative specificity factors remain unknown as well. Approaches that seem promising to reveal these determinants are comparative genetics and pangenomics. With the advent of high-throughput sequencing technologies, it has become possible to analyze bacteria on a whole-genome scale. In 2013, the first complete genome of *S. marcescens* WW4 was sequenced (Chung et al., 2013), and since then, publically available databases are continually updating with new complete genome assemblies. The usage of genomic datasets made it possible to examine particular groups of genomes within species' pangenomes. For example, it allowed characterizing the spread of hospital isolates of *S. marcescens* across the United Kingdom and Ireland (Moradigaravand et al., 2016), identifying antibiotic-resistance factors within historical lineages (Saralegui et al., 2020; Matteoli et al., 2021), and comparing *S. marcescens* with other obligate symbiotic species within the genus (Li et al., 2015; Chen et al., 2017). A recent large-scale pangenomic study enabled defining the population structure of the *Serratia* genomes corresponding to phenotypical traits and ecological niches in which strains resided and showed the relationships between the genetic flow and the emergence of new isolates (Williams et al., 2022).

Despite large progress made over the last few years, pangenomics approaches were not applied to identify putative factors delineating hosts' preference as more attention was paid to virulence rather than specificity. On that account, we reconstructed the *S. marcescens* pangenome using 73 genome assemblies of complete level. In doing so, we compared several popular tools and applied different phylogenetic approaches to develop the most optimal pipeline. Furthermore, we predicted sets of putative specificity factors associated with strains isolated from humans, insects, and plants and carried out functional annotation to characterize functional features of the respective determinants.

Abbreviations: *Sm*, *Serratia marcescens*; HGT, horizontal gene transfer; CYVD, cucurbit yellow vine disease; PGP, plant growth-promoting; SNP, single nucleotide polymorphism; VFDB, virulence factor database; COG, cluster of orthologous genes; GO, gene ontology; CC, cellular component; MF, molecular function; BP, biological process; KEGG, kyoto encyclopedia of genes and genomes; WSS, with-in-sum-of-squares; GIs, genomic islands; ANOVA, analysis of variance; HSD, honestly significant difference; ANI, average nucleotide identity; ML, maximum likelihood; CAMP, cationic antimicrobial peptide; MGEs, mobile genetic elements; pan-GWAS, pan-genome-wide association study; GNAT, Gcn5-related N-Acetyltransferase; aaNAT, Arylalkylamine N-Acetyltransferase; MDR, multidrug resistance.

2. Materials and methods

2.1. Data acquisition

The genome assemblies of *S. marcescens* were downloaded from the NCBI Assembly database (Kitts et al., 2016). Only the genomes deposited in the RefSeq database (O'Leary et al., 2016) with the "complete genome" assembly level remained for further analysis. In order to obtain a high-quality dataset for pangenome analysis, we discarded assemblies with excessively high similarity. We used Assembly-Dereplicator[1] with 99% of identity and a sketch size of 100,000 and selected reference assemblies from clusters accordingly. The respective metadata of the assemblies, including the number of contigs and CDS (coding DNA sequences), genome size, and N50, were obtained using the "ESummary" command from the Entrez v.7.40.20170928+ds-1 package. Next, we calculated the mean GC content of the genomes using a Python v.3.6.9 script implementing the Biopython v1.73 library (Cock et al., 2009). To assess the completeness of the taxonomic markers in the genomes, the BUSCO v5.4.3 package (Simão et al., 2015) was applied with the "*Enterobacteriales*" order specified.

2.2. Pangenome reconstruction

Three popular tools for pangenome were applied, namely, Roary (Page et al., 2015), Panaroo (Tonkin-Hill et al., 2020), and PEPPAN (Zhou et al., 2020). All the programs were launched with default settings and a 95% identity threshold for defining core genes. To check whether the pangenome is open or closed, we transformed gene presence/absence tables into binary matrices and applied the micropan package for the R v.3.6.3 programming language (Snipen and Liland, 2015) to calculate the alpha value based on the Heaps' law. Moreover, we plotted the power-fit curves with a custom R script implementing the ggplot2 library (Wickham, 2016) with 1,000 permutations to calculate the number of genes after expanding the pangenome with new assemblies.

2.3. Phylogenetic analysis

We tested three whole genome-based phylogeny reconstruction approaches to choose the optimal pipeline. First, we picked the trees based on the patterns of the presence/absence of accessory genes. Roary and PEPPAN provided the respective trees as the output. To build the corresponding tree from the Panaroo-reconstructed pangenome, we transformed the gene presence/absence table to pseudoalignment with present genes marked as "C" and absent genes marked as "A" with a custom Python script. The phylogenetic tree was obtained using the FastTree v.2 program with 1,000 bootstrap replications (Price et al., 2010). Second, we obtained core gene alignments. Two aligners were applied when reconstructing pangenomes, namely MAFFT (Katoh and Standley, 2013) and Prank (Löytynoja, 2014). Core SNPs (single nucleotide polymorphisms) were selected with the SNP-sites v2.5.1 tool (Page et al., 2016). We used two types of alignments, namely, concatenated alignments with a single partition specifying the optimal evolutionary model and alignments with the respective individual partitions for each gene cluster. Best-fit evolutionary models were selected using the ModelTest-NG v0.1.7 (Darriba et al., 2020) in the "ml" mode with the best BIC value (Bayesian information criterion). ML (maximum likelihood) phylogenetic trees were built with the RAxML-NG v1.1.0 program (Kozlov et al., 2019) with 1,000 bootstrap replications. Third, we clusterized genomes based on the ANI (average nucleotide identity). Matrix with the pair-wise ANI values was obtained using the Mash v2.3 utility (Ondov et al., 2016) with a k-mer size of 21 and a sketch size of 100,000. The genome clusterization was performed with the "Bclust" function from the shipunov package for R[2] with Euclidean distance metrics and the "complete" clustering method. We then characterized the similarity between the reconstructed trees and assessed the quality of these trees. All the reconstructed trees were topologically compared via the tqDist v1.0.2 library with quartet distance metrics (Sand et al., 2014). Calculated quartet distances were presented as a matrix that was visualized with the ggplot2 package. The quality of the trees was characterized based on two features, namely, mean support of the nodes, and tree balance indices. Three balance indices such as "CollesLike," "Sackin," and "Cophenetic" were calculated using the v2.0 package for R (Mir et al., 2018). The measurements were then plotted on the plane with mean node support on the x-axis and the sum of balance indices divided by the maximum value for each index on the y-axis.

2.4. Virulence and specificity factors analysis

To identify commonly known virulence factors in the analyzed genomes, we used the MMseqs2 tool (Steinegger and Söding, 2017) against the database with protein sequences of virulence factors from VFDB (virulence factor database) (Chen et al., 2005). For each query, the best hit from the database according to the e-value was selected. Next, we retained the hits with not less than 70% query and target coverage coupled with a 70% identity threshold. We then downloaded the metadata of the assemblies from the NCBI BioSample database (Barrett et al., 2012). The genomes were classified according to the infected host, namely, 37 strains were isolated from humans, 4–insects, and 3–plants as well as 29 unassigned strains. Next, binary tables of traits (affected host) were generated using a custom Python script. The data was used to reveal candidate genes associated with host specificity with Scoary v1.6.16 (Brynildsrud et al., 2016). Next, we selected positive associations by calculating the percentage of positive and negative hits using the following formula (1) and then picked only those associations for which the positive ratio was higher than the negative ratio:

$$\frac{pp}{(pp+pn)} > \frac{np}{(nn+np)} \quad (1)$$

where *pp* and *pn* – stand for the number of present and absent genes in genomes attributed to a particular host, respectively, while *np*

1 https://github.com/rrwick/Assembly-Dereplicator

2 https://rdrr.io/cran/shipunov/

and *nn* represent the respective numbers for genomes attributed to other hosts or lacking annotations accordingly.

Scoary-predicted sets of genes were then matched with found virulence factors to identify possible novel virulence determinants. We utilized the gene presence/absence table from the Panaroo-based pangenome and the best ML phylogeny to detect patterns of gene co-occurrence with Coinfinder v1.2.0 (Whelan et al., 2020). Next, tentative host specificity factors were matched with the connected components from the co-occurrence graph.

2.5. Functional annotation

Functional annotation of the protein sequences was carried out with eggNOG standalone tool v2.0.1b-2-g816e190 in the "mmseqs2" search mode (Huerta-Cepas et al., 2017). The attributions belonging to the COG (Cluster of Orthologous Genes) (Tatusov et al., 2000) and GO (Gene Ontology) (Ashburner et al., 2000) annotation systems were examined with the latter including three classes, namely, cellular component (CC), molecular function (MF), and biological process (BP). The annotations of KEGG (Kyoto Encyclopedia of Genes and Genomes) terms were downloaded from the KEGG (Kanehisa and Goto, 2000) site.[3] The text representation was converted to a CSV table using a custom Python script. To characterize functional features of the gene groups (core/accessory, virulence, and host specificity determinants), we applied the custom Python script implementing a hypergeometric test within COG and KEGG annotation systems considering KO codes for the latter. We first build universes for each annotation system with the number of occurrences for each term. Next, we calculated the *p*-values of enrichments related to a certain term for each gene group using the hypergeometric test. The lists of *p*-values were corrected with FDR adjustment (false discovery rate). GO enrichments were calculated utilizing the topGO v.3.15 package (Alexa and Rahnenfuhrer, 2022). We then chose significant enrichments according to adjusted q-values. After that, we performed k-means clustering of significant enrichments. We first obtained a distance matrix based on the pair-wise Szymkiewicz–Simpson (the number of intersections divided by the minimum length of the respective set) and Jaccard coefficients (the number of intersections divided by the length of the union). The optimal number of clusters for the k-means procedure was evaluated with the elbow method (Demidenko, 2018) by manual inspection of depicted with-in-Sum-of-Squares (WSS) numbers. Clustering patterns were subsequently visualized via the "autoplot" function from ggfortify v0.4.11 (Tang et al., 2016).

2.6. Mobile genomic elements search

We described the genomic landscape of *S. marcescens* by identifying mobile genetic elements in the analyzed genomes. Insertions sequences were found using the ISEScan v1.7.2.3 software (Xie and Tang, 2017). To detect prophages, Phigaro v2.3.0 was applied (Starikova et al., 2020). Finally, we revealed genomic islands (GIs) with IslandPath-DIMOB utility (Bertelli and Brinkman, 2018). The CSV tables made were then transformed into coordinates of the respective genetic elements in BED format. Similarly, BED files corresponding to the detected virulence and host-specificity factors were generated using a custom Python script. BED files pertaining to factors and genetic elements were then intersected using the BEDtools "intersect" utility (Quinlan and Hall, 2010) to reveal associations between mobile genetic elements and determinants of virulence and host specificity. We then collected the data on the abundance of different MGEs found in the assemblies attributed to a particular host and performed the ANOVA (analysis of variance) method followed by Tukey's HSD (honestly significant difference) test to reveal statistical differences between genomes.

3. Results

3.1. Genomic properties of the dataset

We first downloaded 99 genome assemblies of *S. marcescens* from the NCBI assembly database (Kitts et al., 2016), and after the dereplication procedure, 73 assemblies remained. The percentage of present markers from the "*Enterobacteriales*" order exceeded 97.3% for all assemblies reaching 99.8% on average (**Supplementary Table 1**). The assemblies possessed similar genome length with the mean value of 5,297,794 b.p. harboring 4,843 CDS with 362 amino acid residues on average (**Supplementary Figures 1A, B**; **Supplementary Table 2**). Notably, the number of hypothetical proteins was proportional to the number of CDS (**Supplementary Figure 1B**) with approximately unannotated 313 proteins lacking per assembly (**Supplementary Table 2**). GC content of the genomes reached 59.5% on average (**Supplementary Figure 1C**; **Supplementary Table 2**). Interestingly, the GC content was inversely proportional to the genome size which also correlated with the number of contigs (**Figure 1A**). This could be explained by plasmids' nucleotide compositions differing from chromosomes (Rocha and Danchin, 2002). The versatile GC content (from 58.5 to 60.2%) corroborated the reported phylogeny-dependent dispersion in GC and negative correlation with plasmids' abundance as well (Ono et al., 2022). Other properties of analyzed genomes were commensurable indicating taxonomical relatedness and the validity of the dataset.

3.2. Assessing pipelines for pangenome analysis

To determine the most optimal pipeline for pangenome reconstruction, we first tested the most popular tools. Panaroo- and PEPPAN-based pangenomes exhibited very high similarity with 15,166 and 15,188 gene clusters (pan-genes) containing 3,700 and 3,600 core genes, while Roary reported a total of 20,380 pan-genes corresponding to 3,341 core genes, respectively (**Figure 1B**). The shape of U-curves reflecting the number of genes present in genomes was comparable for all pangenomes (**Figure 1C**; **Supplementary Table 3**). No considerable internal peaks indicated taxonomical homogeneity between the studied

3 http://geneontology.org/

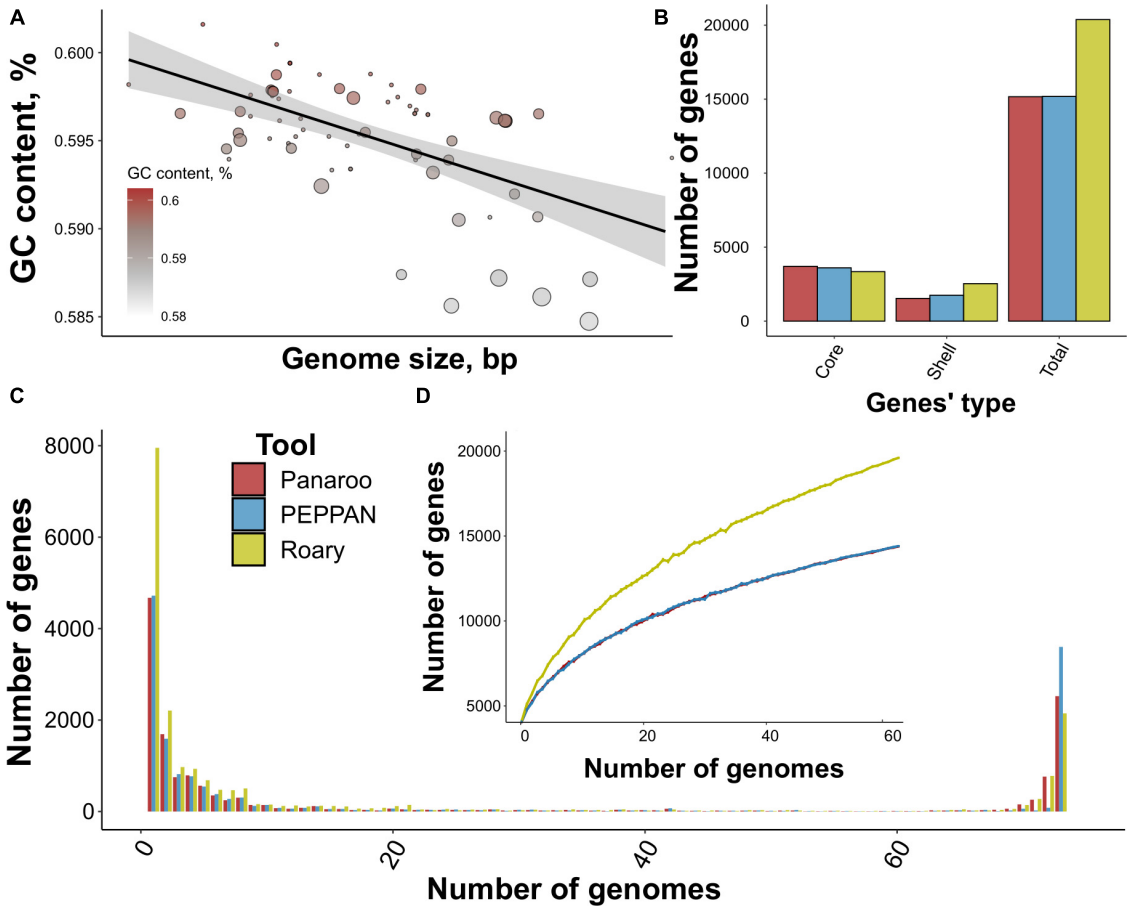


FIGURE 1
Properties of the analyzed dataset comprising *S. marcescens* genome assemblies and main characteristics of the reconstructed pangenome. **(A)** The relation between genome size and mean GC content. Dots' size represents the number of contigs per assembly, and the intensity of the color codes GC content. A linear relationship between the properties is illustrated by a black line. **(B)** The number of gene groups (core, shell, and total) within pangenomes built using three tools, namely Panaroo, PEPPAN, and Roary. **(C)** U-curves of the respective pangenomes showing the frequency of pan-genes presented in a certain number of genomes. **(D)** The power-fit curve of *S. marcescens* pangenomes based on 1,000 permutations of a random sampling of genomes. Bars denote standard deviation.

assemblies. Still, some remarkable differences were detected. First, the peak corresponding to gene clusters presented in all 73 genomes was the highest for PEPPAN-built pangenome (3,600 pan-genes vs. 2,523 for Panaroo). Second, Roary provided a considerably larger number of unique genes of 7,954 in comparison with Panaroo and PEPPAN reporting 4,674 and 4,720 genes, respectively.

All three power-fit curves did not reach saturation thus indicating that *S. marcescens* pangenome could be considered open. However, Roary provided a curve resembling linear function probably due to the inflation of unique genes (**Figure 1D**). We also calculated the α parameters according to Heaps' law. The estimates spanned from 0.52 to 0.64 further proving the pangenome to be open as $\alpha < 1$ implies openness (Tettelin et al., 2008).

In general, all the pangenomic approaches provided comparable results, however, PEPPAN and Panaroo provided a lesser number of accessory genes, thus minimizing their inflation which could lead to over-estimation of pangenome openness. However, PEPPAN's limitation lies in the absence of pan-genes' alignments making it impossible to build full-genome phylogeny. Therefore, we proposed Panaroo to be the most suitable tool for reconstructing the *S. marcescens* pangenome and, hence, considered the respective pangenome as the reference for further examination.

3.3. Core and accessory genes-based phylogenies

In addition to pangenome analysis, we also compared several phylogenomic pipelines. We considered three main approaches: using core gene alignment representing relationships between lineages (Ding et al., 2018), clustering accessory genes thus reflecting gene gain/loss and HGT (Whelan et al., 2021), and calculating average nucleotide identity (ANI) (**Supplementary Figure 2**) thought to reflect evolutionary history depending on overall genetic distance (Gosselin et al., 2022). Moreover, we either applied partition schemes for each core gene or utilized a single model.

We found that all the trees showed considerable topological similarity with a mean value of 92%. Importantly, phylogenies based on core genes were highly congruent with topological similarity exceeding 96% (**Supplementary Figure 3A**;

Supplementary Table 4). On the other hand, clusterization-based trees were less consistent (Supplementary Figure 2A). Genome grouping according to ANI resembled core genome phylogenies more than those based on accessory genes' distributions. In general, the highest mean similarity (94.5%) with all trees was shown for the phylogeny reconstructed on the basis of core genes obtained from the Panaroo-reconstructed pangenome aligned with MAFFT without partitions (Supplementary Figure 3B; Supplementary Table 4).

We also compared the quality of the reconstructed trees considering considered two measures, namely, mean branch support and tree balance. As it is evident from the graph, clusterizations gained higher support values than true core gene phylogenies and were more balanced as well (Supplementary Figure 3C; Supplementary Table 5). This observation, however, could be explained by less sophisticated computations applied to obtain such trees. The highest branch support was reported for the aforementioned phylogeny with the highest mean similarity, albeit being less balanced. Nevertheless, alignment-based phylogenies differed negligibly in terms of balance. According to our results, partitioning slightly increases tree balance but not branch support (Supplementary Figure 3C). Given that the former is not indicative of quality (Stam, 2002), we believe that an almost unnoticeable gain in balance does not cost such a considerable rise in computational cost (Supplementary Table 6). Taking into account all the facts mentioned, we selected a best-choice pipeline involving Panaroo for pangenome reconstruction and unpartitioned core gene alignment made by MAFFT for building a reference tree.

3.4. Distribution of host specificity and virulence factors

To reveal candidate specificity factors of *S. marcescens*, we attributed genomes to affected hosts specified in the BioSample database and found three main host groups, namely, humans, insects, and plants, with which 37, 4, and 3 assemblies were associated, respectively (Supplementary Table 7). The remaining 29 assemblies represented primarily environmental strains with unknown host specificity. We found that 4,852 core gene clusters were shared between all the groups (Figure 2A), whereas non-shared pan-genes reached 3,618, 241, and 273 for isolates infecting humans, plants, and insects, respectively. Interestingly, genomes related to human infection included more common gene clusters with those infesting plants than with insect-infecting (122 vs. 146), while the latter two groups shared 15 pan-genes only.

Given the unequal number of host-wise genomes, we selected only positive associations (see Section "Materials and methods"). Due to the scarcity of data, only 8 pan-genes related to human infection regarding Benjamini-Hochberg-corrected *p*-values we reported. These genes included surface lipoproteins and transporters (Supplementary Table 8). We, however, aimed to reveal general functional features of tentative specificity factors and thus selected all found hits based on raw *p*-values (Supplementary Table 9). There were only 3 reported gene clusters common for insect and plant infestations (Supplementary Figure 4A). When disclosing gene clusters and considering the existence of genes within pangenome clusters in host-attributed assemblies, we found that 227 gene clusters were common for all three groups, while 116, 4, and 241 pan-genes were unique for assemblies associated with humans, insects, and plants, respectively (Figure 2B). More specificity-related pan-genes were common for human- and insect-associated strains than for those infecting humans and plants (102 vs. 72), whereas groups attributed to plants and insects lacked shared clusters. After that, we analyzed the abundance of selected specificity markers among the genomes (Figure 2C). Plant-related gene clusters showed the lowest frequency with four genome assemblies on average compared with 9 and 24 genomes for putative specificity factors related to insect and human infections. In addition to this, plant-associated clusters contained 35% of hypothetical proteins in contrast to insect- and human-attributed clusters with 17 and 19%, respectively (Supplementary Figure 4B).

Next, we carried out the screening of known virulence factors from the VFDB (virulence factor database) among the analyzed genomes to find possible relations between virulence and specificity. We identified 7,955 virulence factors belonging to 231 pan-genomic clusters. On average, homologs of virulence factors were presented in 67 genomes, and the vast majority of factors represented core genes with a total of 6,714 (Figure 2D). The homologs showed 81% mean identity with the respective references from the database. Interestingly, three main groups of virulence genes in the context of the identity were found (Supplementary Figure 4C). They included core and shell genes with a mean identity of 80%, higher than 90%, or not exceeding 80%. Remarkably, a small fraction of virulence factors corresponded to specificity determinants. Only 81, 8, and 9 virulence genes constituted 1.5, 3.8, and 1.9% of specificity factors associated with humans, insects, and plants, respectively (Figure 2E).

We then analyzed the phylogeny-wise distribution of these specificity factors (Figures 3A, B; Supplementary Table 10), and selected seven compact phylogenetic subclades forming distinct patterns in terms of specificity factors' abundance (Figure 3A). These clades are also visible, albeit not explicitly when examining the distribution of all accessory genes (Supplementary Figures 4D, E). However, the dependence between host-wise annotation and phylogeny was not straightforward (Figure 3C). For instance, clade 2 encompassed primarily environmental samples, including two genomes isolated from insects and two strains found in plants, and was enriched with the respective specificity factors, yet, including one clinical isolate. Clades 1 and 7 contained primarily isolates lacking host annotations and were poor with specificity factors, and this was especially notable for clade 1 (Figures 3A, B). Clades 3 and 4–6 were presented mostly by human-associated strains. In spite of this, clade 4 encompassed one strain residing on plants, whereas clade 5 – insect-infecting isolate. Notably, the respective distribution of virulence determinants was uniform and independent from the phylogeny (Figure 3D).

After that, we used the CoinFinder tool to dissect the co-occurrence of specificity factors. The co-occurrence graph contained 12 connected components with the first harboring soft-core genes and the others containing shell genes with 59 assemblies minimum (80% of presence). Only one human-related factor (sugar efflux transporter) formed a connected component with RNA chaperone/antiterminator CspA and isochorismatase family protein (Supplementary Table 11). The absence of host-wise co-occurrence networks could be caused by a limited number of genomes; thus further large-scale studies may reveal more

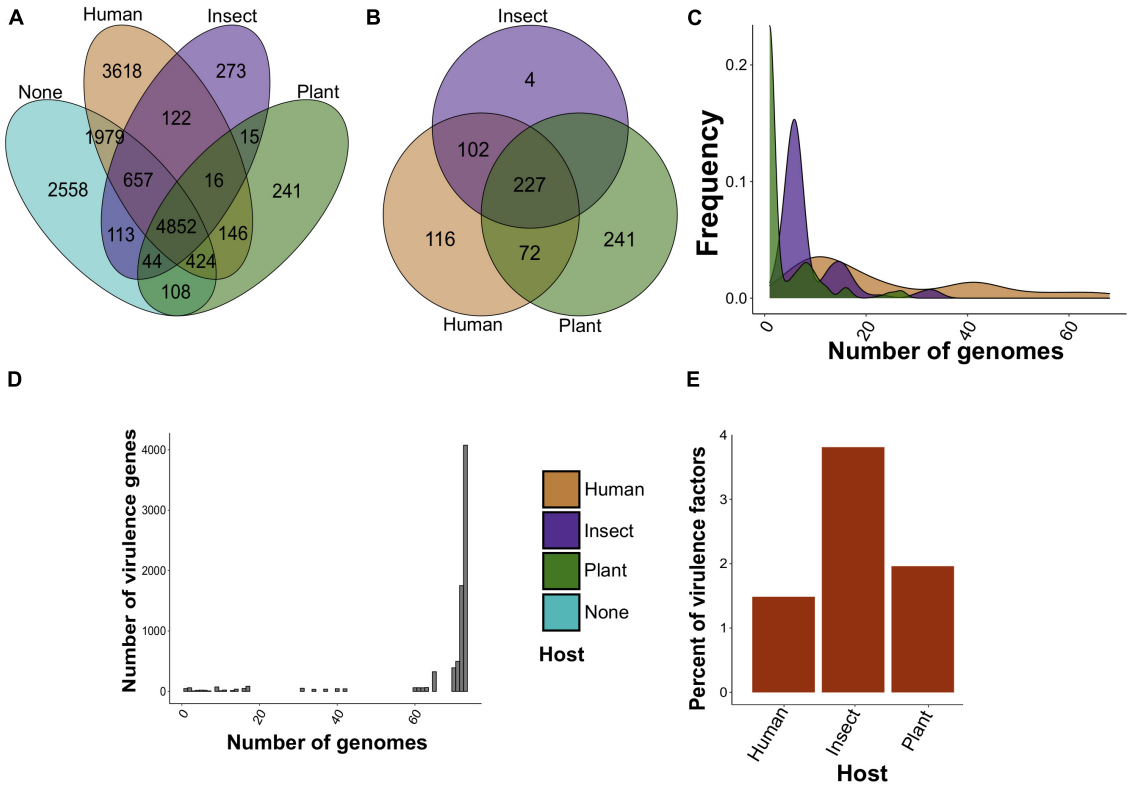


FIGURE 2
The number of found specificity and virulence factors in *S. marcescens* genome assemblies. **(A)** Venn diagram of common and non-shared gene clusters between assemblies attributed to a particular host (human, insect, and plant). Clusters are considered common if at least one gene from the orthologous cluster is shared by assemblies isolated from different hosts. "None" indicates the absence of the host in the metadata. **(B)** The same results for pangenome clusters reported as specificity factors using Scoary. **(C)** Frequency distribution of specificity factors. Plotted on the y-axis is the number of genomes in which the respective factor is present. **(D)** The abundance of virulence genes. The y-axis reflects the frequency of virulence factors possessed by a certain number of genomes. **(E)** Percentage of specificity factors having a respective homolog in the VFDB.

associations. Furthermore, this observation might also be explained by the richness of dispensable pangenome components.

3.5. Functional annotation of gene clusters

To reveal functional features of gene groups, we performed group-wise enrichment analysis and compared pangenome orthologous clusters (core, accessory, and unique) and virulence determinants (core and accessory) with host-wise specificity factors. On the whole, the number of annotations, for each group, was equal to the number of genes in the clusters, except for those containing paralogs which gained a higher number of functional terms (Supplementary Figure 5). There were 85 clusters in which some genes were annotated with others having no annotation terms (Supplementary Table 12). When considering the absolute number of terms per gene cluster, several distinct lines were found corresponding to the mean number of annotation terms per gene (Supplementary Figure 6). However, within GO (Gene ontology) system, a more dispersed distribution with high variance for core genes was found (Supplementary Figure 6). Notably, 99.8% of gene clusters shared identical sets of functional terms for each gene within the cluster (Supplementary Figure 7). Those that were non-identical mostly lacked annotations for certain genes. Core genes, both pan-genomic and core virulence, contained more clusters with functional terms, especially when considering COG (Cluster of Orthologous Genes) and KEGG (Kyoto Encyclopedia of Genes and Genomes) systems (Supplementary Figure 8). Human-associated specificity factors incorporated slightly more annotated clusters than those related to plant and insect infections, respectively (Supplementary Figure 8). Interestingly, clusters with incomplete annotations were likely involved in processes affecting virulence, e.g., cell motility, ion binding, two-component system activity, enteric infections, etc. (Supplementary Figures 9A–E; Supplementary Table 13). We thus might assume that missed data in the existing databases could be supplemented using terms from pangenome clusters.

Within the COG system, core genes were enriched with primary cellular and metabolically processes' categories, while accessory genes, in their turn, were linked with extracellular structures, secondary metabolites' synthesis, motility, cell wall biogenesis, replication and recombination, and defense mechanisms, which was a characteristic of virulence determinants (both core and accessory) as well (Figure 4A; Supplementary Tables 14, 15). Human-associated genes incorporated categories linked with transcription and cell cycle control/division. Infecting insects was related to intracellular trafficking and secretion, whereas preference

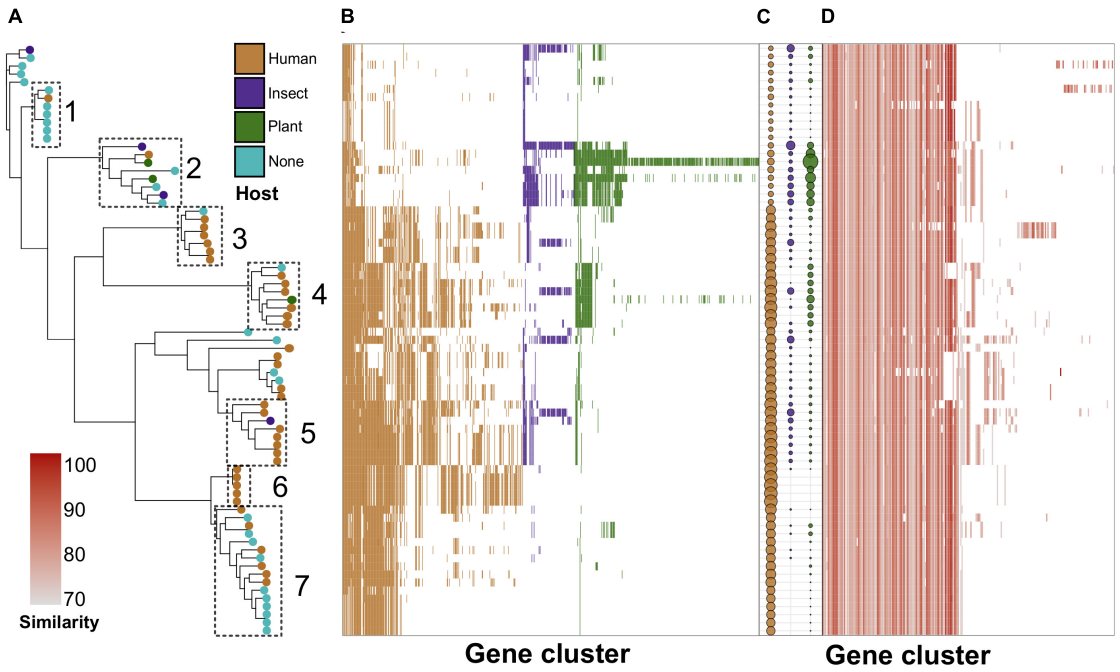


FIGURE 3
Phylogeny-wise distribution of specificity determinants and known virulence factors in *S. marcescens* genomes. **(A)** Reference ML (maximum likelihood) tree with leaves colored by the host. Dotted rectangles denote selected subclades with corresponding numbers assigned to them. **(B)** The presence/absence of specificity factors among the genomes in the heatmap ordered by the adjacent reference tree. **(C)** The number of host-wise specificity factors in a particular genome. The size of the dots is proportional to the number of respective genes. **(D)** Distribution of virulence factors in the genomic dataset. The intensity of the color denotes sequence similarity with the closest homolog from the VFDB (virulence factor database).

toward plants implied the same enrichments as those reported for unique genes, i.e., replication and recombination, and defense mechanisms.

Over-representation of KEGG terms was consistent with the COG enrichments showing core genes to participate in base cell and metabolic processes (glycolysis, transport, translation, etc.), and accessory component to antibiotic resistance, secretion, glycerolipid metabolism, and, surprisingly, the plant-pathogen interaction which was present in human-associated but not plant-related gene set (Figure 4B). Unique genes show pathways related to enteric infections and invasion of epithelial cells. The gene group delineating infection in humans was functionally similar virulence factors, such as biofilm formation, ABC transporters, two-component system, lipopolysaccharide biosynthesis, cationic antimicrobial peptide (CAMP) resistance, secretion, and polyketide synthesis. Other annotations related to this set of genes were the phosphotransferase system and resistance to beta-lactams. The main functional hallmarks of assemblies isolated from insects are connected to secondary metabolites' synthesis, including terpenoids, polyketides, and phenylpropanoids. Apart from terms within the accessory genes, samples collected from plants were enriched with the degradation of chemical substances such as xylene, toluene, and fluorobenzoate.

Speaking of predicted localization from the GO cellular component system, core, as well as core virulence genes' products, are localized in intracellular compartments, while proteins encoded by accessory genes, and specificity factors associated with humans and insects reside in cellular membranes, periplasmic space, and cell periphery (Figure 4C). The group of genes determining plant infection shared the same terms as singletons (extracellular capsule and outer membrane). Enrichments within the biological processes category for general pangenome components and virulence genes all pertained to different primary cellular and metabolic pathways, including nitrogen compounds carbohydrate metabolism, transport transcription, and regulation (Figure 5A). In contrast, determinants of preference toward hosts displayed more specific annotations as follows: cofactor metabolic processes (humans), oxidation-reduction, organic cyclic, and heterocyclic compounds metabolism (insects), interspecies interaction, pathogenesis, response to biotic stimulus, filamentous growth, and glucosamine metabolism (plants). Fitting into the general frame, molecular functions of the products encoded by core genes encompassed binging of various chemical moieties including proteins, ions, and organic compounds, while the accessory component and known virulence determinants are engaged in the transport and transferring of phosphorous groups and DNA binging thus implying their possible role in signal transduction and transcriptional regulation (Figure 5B). In addition to terms shared with core genes and virulence factors (binding and transferase activity), group-specific characteristics of specificity determinants linked with human infection included phosphoryl transfer-driven membrane transport. Molecular functions of genes attributed to insect hosts represented a spectrum of metabolic reactions governed by diverse enzymes such as succinyltransferase, aldo-keto reductase, 5-aminolevulinate synthase, and others. Finally, over-represented terms encompassed by plant-associated genes were porins and channels, deaminase, and oxidoreductase activities.

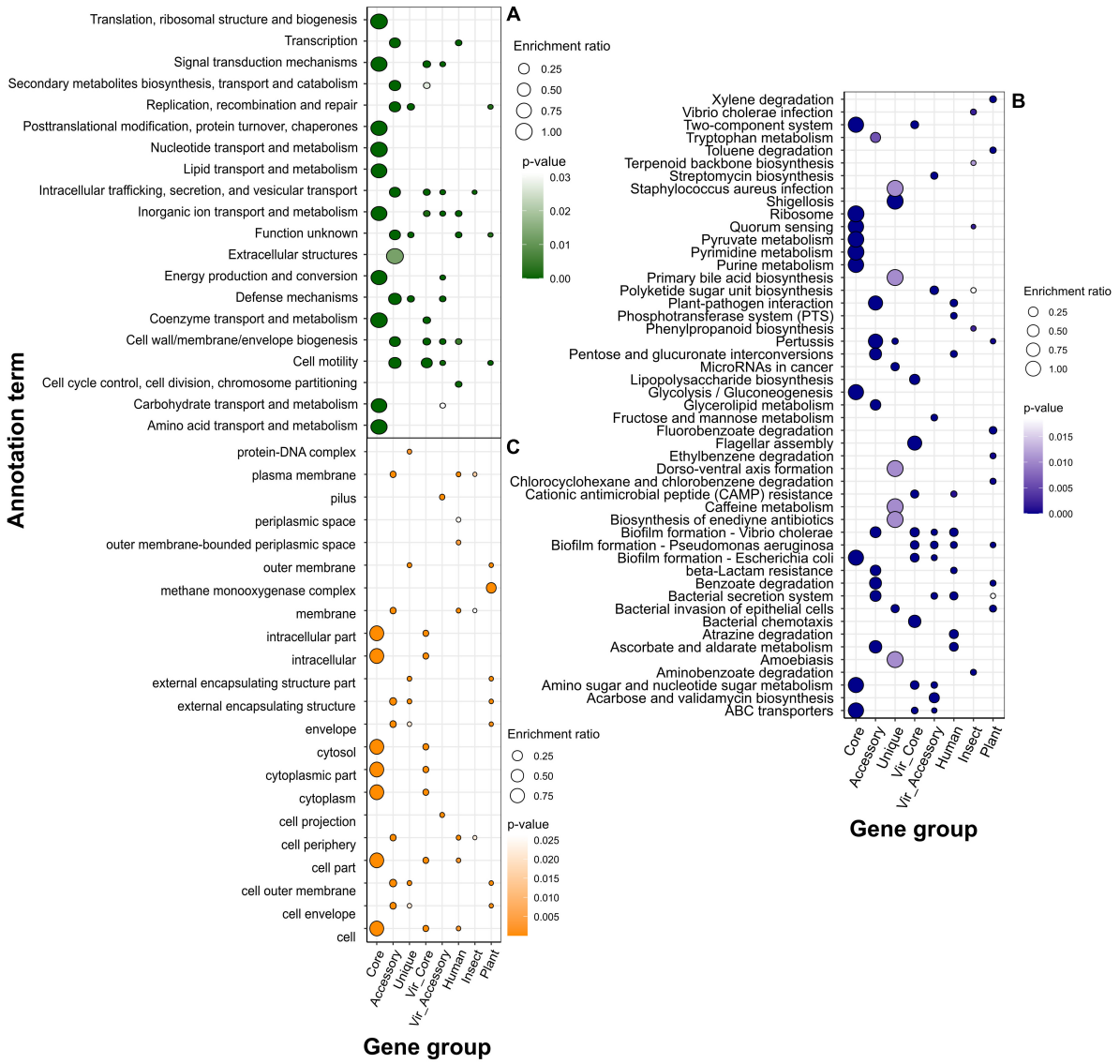


FIGURE 4
Over-represented annotation terms in different gene groups, namely, pangenomic (core, accessory, and unique), virulence (core and accessory), and specificity genes attributed to a particular host (human, insect, plant) using COG (Cluster of Orthologous Genes) **(A)**, KEGG (Kyoto Encyclopedia of Genes and Genomes) **(B)**, and GO (Gene Ontology) cellular component **(C)** annotation systems. The color denotes adjusted *p*-values, and the dot size depicts the enrichment ratio.

With the aim to find functional similarity between host specificity factors, we used all significant enrichments of six gene groups: three host-wise accessory genes and Scoary-reported associations by performing the k-means clustering procedure (Supplementary Table 16). The elbow method revealed that four clusters were optimal to be used (Supplementary Figures 10A–D). All clusterizations were convergent and provided the same grouping patterns (Supplementary Figures 11A–D). The largest cluster united host-wise accessory components, whilst human-, insect-, and plant-associated specificity determinants fell into distinct clusters. It thus could be proposed that genomes of strains isolated from different hosts are functionally similar *per se*, however, specificity determinants show varying functional pathways.

3.6. Mobile genetic elements associated with virulence specificity

To find possible connections between virulence and specificity with mobile genetic elements (MGEs) we searched for these elements in the genomic dataset used. We considered three types of MGEs: insertion sequences (ISs), prophages, and genetic islands (GIs). We found a positive correlation between the number of the elements, however, significant associations according to the Pearson test were reported for prophages vs. GIs (p-value < 6.5e-06) and ISs vs. GIs (p-value < 2.2e-16) with the corresponding correlation coefficients of 0.5 and 0.8 (Figures 6A–C; Supplementary Table 17). Median values of MGEs' abundance reached 10, 3, and 7 for ISs, prophages, and GIs, respectively.

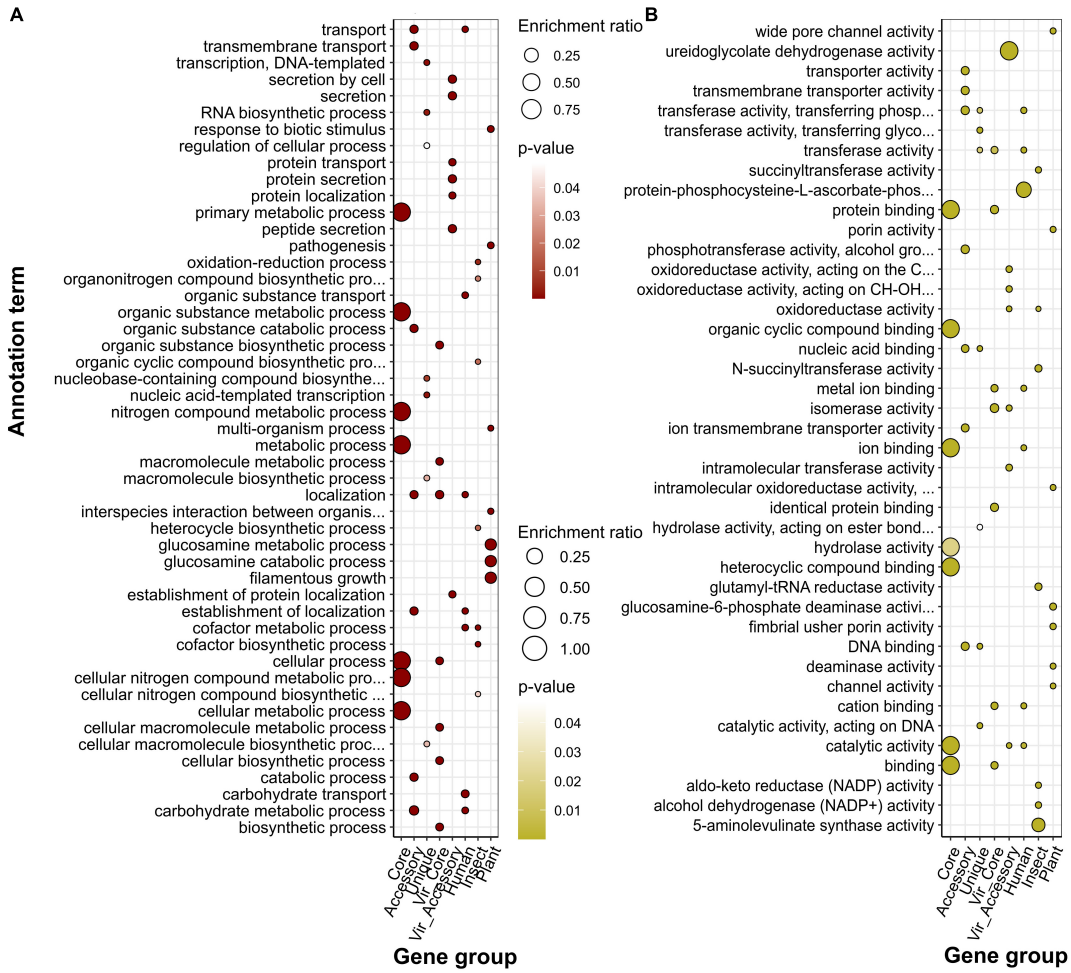


FIGURE 5
Over-represented annotation terms in different gene groups, namely, pangenomic (core, accessory, and unique), virulence (core and accessory), and specificity genes attributed to a particular host (human, insect, plant) using GO (Gene Ontology) biological processes **(A)** and molecular function **(B)** ontologies. The color denotes adjusted *p*-values, and the dot size depicts the enrichment ratio.

When considering the assemblies grouped by the attributed host, we found that genomes isolated from plants and insects bore half as many ISs as other assemblies, while the mean frequency of prophages and GIs were comparable and close to overall median estimates (**Figures 6D–F**).

Having shown no remarkable difference between the absolute number of the MGEs, we decided to evaluate the percentage of genes from the above-defined gene groups whose genomic coordinates intersected with the elements (**Figures 6G–I**). Of note, the smallest percentage was shown for core and virulence genes with the exception of accessory virulence factors with circa 10% of genes located within prophages. Accessory and especially unique genes were most prominently linked with MGEs with almost 60% of singletons residing in GIs. Insertions contained the lowest number of genes. Nevertheless, 2% of human-associated genes and 4% of singletons were located within insertions. The most conspicuous feature of specificity factors attributed to insect and plant infections was a high proportion (more than 40%) of genetic islands harboring them. Next, we verified how many of the host-wise specificity genes were located within MGEs (**Supplementary Table 18**). Human-attributed assemblies contained 96 insertions bearing 1 to 2 genes. Furthermore, we found 13 prophages with more than 18 factors in human-infecting strains. Two assemblies isolated from plants were enriched with genomic islands. One of these (GCF_001417865.2) harbored tree GIs with 69, 37, and 28 specificity determinants, and the other (GCF_011602465.1)–one island bearing 15 genes. Finally, we identified two genomic islands with 41 insect-associated genes both in two assemblies, namely GCF_013122155.1 and GCF_015160915.1.

4. Discussion

Due to the severe impact of *S. marcescens*-induced nosocomial infections, current studies focus on revealing phylogenetic lineages of clinical isolates and/or identifying antibiotic-resistance genes (Moradigaravand et al., 2016; Ono et al., 2022). By contrast, our work aims to explore *S. marcescens* adaptations to certain hosts, namely, humans, insects, and plants using a pangenome-wide associating method (pan-GWAS) which was applied by Ono et al. (2022) when revealing accessory genes specifically conserved in two large clinical and hospital-associated *S. marcescens* clades.

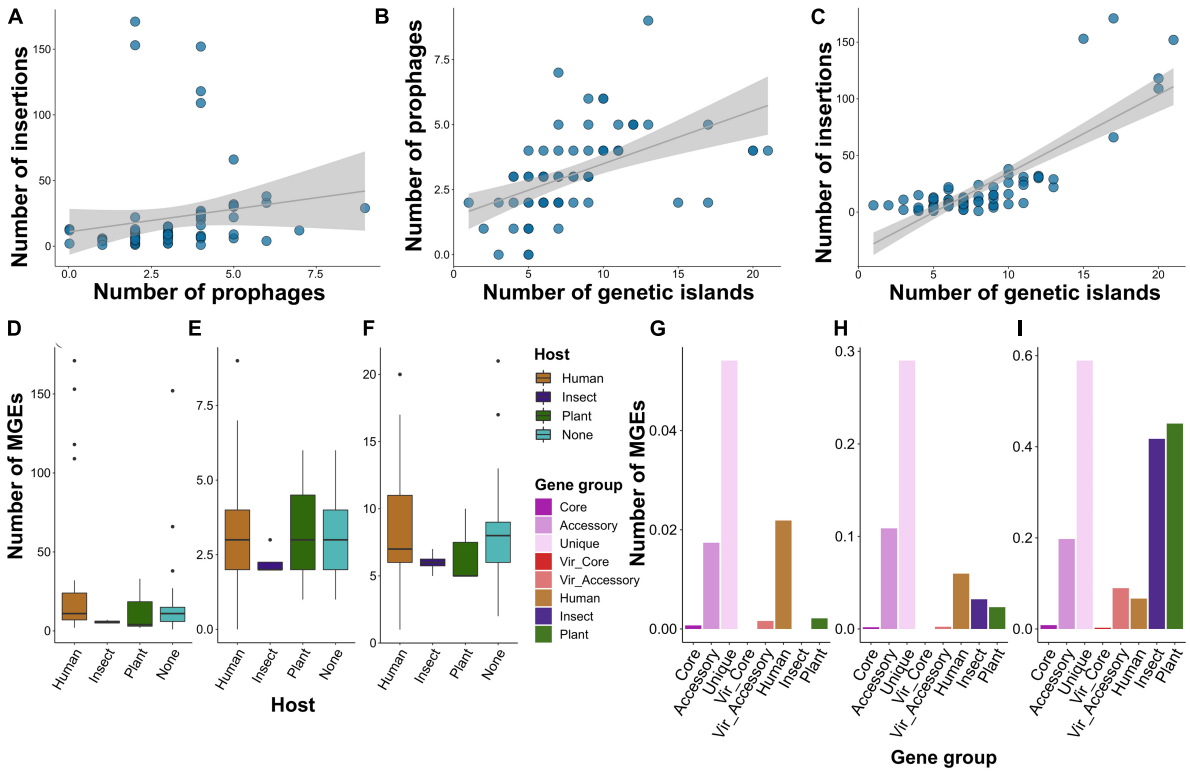


FIGURE 6
Shown are the features of mobile genetic elements (MGEs) presented in *S. marcescens* genomes. **(A)** Dependence between the number of insertions and phages, genetic islands and phages **(B)**, insertions and genetic islands **(C)**. Each dot represents an assembly, and the gray line denotes linear relationships between the abundance of MGEs. **(D)** The number of insertions, phages **(E)**, and genetic islands **(F)** among *S. marcescens* genomes grouped by the host attributed to the respective assemblies. **(G)** The percentage of genes within gene groups, namely, pangenomic (core, accessory, and unique), virulence (core and accessory), and specificity genes attributed to a particular host (human, insect, plant), located within the sequences of insertions, phages **(H)**, and genetic islands **(I)**.

Nonetheless, to the best of our knowledge, a host-wise pangenomic exploration of putative specificity factors has not been done yet. Abreo and Altier (2019) conducted the research with relatively similar goals regarding plant growth promoting (PGP) and virulence traits, yet the authors used a pre-existing list of virulence (22) and PGP (21) genes, whilst concerning that this approach could leave unknown genetic traits hidden within the genomes.

We analyzed 73 *S. marcescens* genomes with a complete level of assembly and performed a methodological comparison to choose the most optimal pipeline. Circa 76% of the species' pangenome belonged to the accessory (**Figure 2C**), which matched with other topical research items in which comparable values were obtained (Moradigaravand et al., 2016; Abreo and Altier, 2019; Ono et al., 2022; Williams et al., 2022). These inferences coupled with an open pangenome, first explicitly estimated by us, imply considerable genetic variability of *S. marcescens* strains and, therefore, might explain such a broad host range and multiple activities. Another indication of genetic plasticity was a high proportion of mobile diverse MGEs bearing accessory genes (**Figures 6G–I**). It is worth noting, that the abundance of MGEs *per se* is not indicative of the preferred host (**Figures 6D–F**). That being said, human-isolated samples contained plenty of prophages, and the list of specificity factors included many phage tail and baseplate assembly proteins. That is of importance as many phages associated with virulent strains encode extracellular toxins, effector proteins participating in invasion (Fortier and Sekulovic, 2013). Plant- and insect-associated factors, by contrast, tend to be located on GIs. Their role in shaping the genome of *S. marcescens* was linked to the emergence of MDR (multidrug-resistant) phenotypes (Mataseje et al., 2014; Gambino et al., 2021). Of particular note, *Serratia* sp. SCBI was found to harbor unique GI related to enteric infections of *Caenorhabditis briggsae* (Abebe-Akele et al., 2015). Henceforth, it is not the quantity but the quality of MGEs that probably delineates adaptation to a particular host.

We identified eight human-associated specificity factors with significant adjusted *p*-value. Of these, MgtC/SapB was found to induce biofilm formation of *Pseudomonas aeruginosa* (Redfern et al., 2021). The YgdI/YgdR lipoprotein was also found in the biofilm matrix proteome from a *P. aeruginosa* clinical isolate (Egorova et al., 2022). LysR family proteins are recognized as well-known regulators of multiple *S. marcescens* activities, including cell motility, exopolysaccharide synthesis, and biofilm formation (Pan et al., 2020). On top of that, these transcription factors account for the increased virulence of bacterial pathogens, e.g., *P. aeruginosa* (Yeung et al., 2009) and *Vibrio cholera* (Bogard et al., 2012). Finally, examinations of diverse bacterial pathogens' virulence mechanisms revealed the essential role of ABC transporters (Garmory and Titball, 2004). Among other proteins encoded by genes related to human infection according to empirical *p*-values, we identified ankyrin repeat domain-containing protein which could possibly

manipulate hosts functions (Al-Khodor et al., 2010), and porins which are considered effectors of infection (Achouak et al., 2001). Top phenome-wide inferences successfully collate with the most notable functional enrichments such as transport, including iron acquisition, and periplasmic space. Periplasmic proteins (Liu et al., 2009; Moreira et al., 2013) and the process of ion acquisition (Payne, 1993; Rodriguez and Smith, 2006) are considered essential in managing bacterial pathogenesis.

Three top-scored putative specificity factors linked with insect infections were GNAT (Gcn5-related N-acetyltransferase), cysteine hydrolase, and isochorismatase, corroborating common functional traits attributed to metabolic pathways, namely, oxidases, reductases, transferases, and dehydrogenases. GNAT proteins acetylate different substrates and play a role in bacterial antibiotic resistance (Czub et al., 2018). Arylalkylamine N-acetyltransferase (aaNAT) belonging to the GNAT family has received attention due to its involvement in regulating the synthesis of neuromediators in insects (Tang et al., 2022). Hence, it could be hypothesized *S. marcescens* pathogenesis in insects may well be carried out via neurotoxicity. Zoopathogenic bacteria also utilize cysteine hydrolase to impair host immune responses (Kędzior et al., 2016). Another mechanism to modulate the behavior of host cells lies in triggering the autophagy of host cells provoked by isochorismatase activity (Wang et al., 2016).

Selected putative contributors to host preference toward plants fell into two categories: virulence and PGP agents. The first category encompassed the MurR/RpiR family transcriptional regulator, the homolog of which was demonstrated to regulate cell wall synthesis (Jaeger and Mayer, 2008). It is quite probable that these alterations can lead to evading plant recognition and immunity via the reduction of host perception of cell wall peptidoglycans (Gust, 2015). We also found multiple pan-GWAS signals corresponding to the type VI secretion system, a key infection mediator of plant pathogens (Rainey, 1999). The second category comprised oxalurate catabolism protein HpxZ and amino acid ABC transporter permease. Oxalurate metabolism constitutes a branched complex network of chemical transformations ending in a release of ammonium ions, which alleviate plants' tolerance during nitrogen deficiency (Izaguirre-Mayoral et al., 2018). As for amino acid transporters, they were shown to participate in plants' dialogue with rhizospheric bacteria, e.g., *Pseudomonas fluorescens* (Bernal et al., 2018). Like with other hosts, plant-associated specificity determinants corresponded to a functional annotation with enrichments related to porins, cell wall and encapsulating structures, and interspecies interactions. The functional feature that deserves particular attention is xenobiotic catabolism. Being probiotic bacteria, some strains of *S. marcescens* ensure plant resistance to chemical pollutants due to their capability of neutralizing benzo(a)pyrene (Kotoky and Pandey, 2020), fluorobenzoate (Zhang et al., 2020), toluene and other toxic moieties (Stancu, 2016).

We also examined how the revealed specificity factors are distributed within the genomes analyzed regarding phylogenetic relationships between them (**Figures 3D, E**). Of the subclades selected, clade 2 represented four strains reported to be pathogenic against insects and plants, respectively. This clade was poor in human-associated factors but enriched with those attributed to plants and insects. The clade contained two entomopathogenic isolates, namely strain FY and Byron invading intestines *Drosophila melanogaster* (Liu et al., 2020), and *Curculio caryae* (Wu et al., 2021), respectively. Several strains constituting this clade were included in the analysis performed by Abreo and Altier (2019) as a part of the environmental clade named 1c encompassing one strain with PGP properties and lacking clinical isolates. This strain was absent in our analysis as we collected only complete level assemblies. However, the clade contained one presumably plant-pathogenic sample B3R3 shown to be close to the *S. marcescens* strain causing leaf spot disease on industrial hemp (*Cannabis sativa* L.) (Schappe et al., 2019) and BP2 strain found in seeds of *Jatropha curcas* with unknown activity. It must be noted that strain U36365 which fell into this phylogenetic group was described as resistant to antibiotic therapy causative agent of urinary tract infection (Sahni et al., 2016). Inexplicably, the isolate formed green-colored but not red-pigmented colonies (Sahni et al., 2016). Clades 3 and 6 contained almost exclusively human-infecting clinical samples, whereas clades 1 and 7 were presented by infectious to human strains coupled with assemblies of unknown origin. Clades 4 and 5 were of intriguing composition. Even though they contained primarily human-associated genomes, PGPB strain RSC-14 pertained to clade 4 (Khan et al., 2017), and strain N10A28, claimed to be a pathogen of *Apis mellifera*, – to clade 5. Another remarkable observation that we found was the distribution of insect-related specificity factors along clades from 2 to 5, inclusively, since 7 assemblies were enriched with these determinants comparable with those infecting insects. To this set of genomes belonged 2 isolates of environmental origin, namely rhizosphere soil and water surface of an oligotrophic pond, seed-residing BP2 strain, three human pathogens, and one genome of unknown origin. It hence may be proposed that these genomes may possess hidden insecticidal activity.

We should note that a relatively high portion of strains not attributed to any host has led to several important caveats, decreasing the number of statistically significant associations. Moreover, there is an obvious skew toward studies related to pathogenic clinical isolates which could lower the statistical power of the inferences. We deliberately included assemblies with unassigned hosts, thereby making the sample as close as possible to the general population instead of cherry-picking based on predetermined properties of interest. Nonetheless, the methodology used provided us with certain novel genetic determinants, underlining its high sensitivity.

5. Conclusion

Although belonging to the *Enterobacteriaceae* family, *S. marcescens* is able to colonize not only the mammalian intestine but a wide range of hosts, including invertebrates and even plants. Such specialization to different ecological niches raises the question regarding the diversity of molecular factors, which determine the host specificity. Having studied the pangenome of a relatively limited but high-quality and complete set of genomes, we have found significant variability in virulence factors, with only housekeeping genes forming the core part. By analyzing the reconstructed pangenome, we identified novel factors that possibly determine the adaptation to the particular group of

hosts. Of note, patterns of functional enrichments allowed us to hypothesize that the respective host preference is carried out through distinct molecular mechanisms of virulence. Moreover, the found candidates only scarcely intersect with the known virulence factors presented in the VFDB resource. Not only do specific genes delineate host adaptation but also *S. marcescens* isolates attributed to different hosts harbor group-specific mobile elements, which provides insights into possible ways of how the specificity factors are spread through bacterial populations. To sum up, our methodology helped us to reveal new factors delineating host specificity on the basis of the relatively small dataset with scarce metadata, which is quite common in such studies. The found incongruence between the distributions and the abundance of host specificity determinants and known virulence factors might imply that virulence itself does not delineate adaptations to particular hosts.

Data availability statement

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

Author contributions

AS performed pangenome and phylogeny reconstruction, created scripts, visualized the results, and wrote and revised the original manuscript. AM performed pan-GWAS and validated the results. IS contributed to text editing. AN contributed to text editing and writing, project administration, and funding. KA conceptualized and supervised the study and wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

The manuscript was made with the support of the Ministry of Science and Higher Education of the Russian Federation in accordance with agreement No 075-15-2021-1055 date September 28, 2021 on providing a grant in the form of subsidies from the Federal budget of the Russian Federation. The grant was provided for the implementation of the project: "Mobilization of the genetic resources of microorganisms on the basis of the Russian Collection of Agricultural Microorganisms (RCAM) at the All-Russia Research Institute for Agricultural Microbiology (ARRIAM) according to the network principle of organization."

Acknowledgments

We authors thank Oksana Belousova for English editing.

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.2023.1211999/full#supplementary-material

References

Abebe-Akele, F., Tisa, L. S., Cooper, V. S., Hatcher, P. J., Abebe, E., and Thomas, W. K. (2015). Genome sequence and comparative analysis of a putative entomopathogenic *Serratia* isolated from *Caenorhabditis briggsae*. *BMC Genomics* 16:531. doi: 10.1186/s12864-015-1697-8

Abreo, E., and Altier, N. (2019). Pangenome of *Serratia marcescens* strains from nosocomial and environmental origins reveals different populations and the links between them. *Sci. Rep.* 9:46. doi: 10.1038/s41598-018-37118-0

Achouak, W., Heulin, T., and Pagès, J.-M. (2001). Multiple facets of bacterial porins. *FEMS Microbiol. Lett.* 199, 1–7. doi: 10.1111/j.1574-6968.2001.tb10642.x

Alexa, A., and Rahnenfuhrer, J. (2022). *topGO: Enrichment analysis for gene ontology. R package version 2.48.0*.

Al-Khodor, S., Price, C. T., Kalia, A., and Abu Kwaik, Y. (2010). Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol.* 18, 132–139. doi: 10.1016/j.tim.2009.11.004

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Res.* 40, D57–D63. doi: 10.1093/nar/gkr1163

Baya, A. M., Toranzo, A. E., Lupiani, B., Santos, Y., and Hetrick, F. M. (1992). *Serratia marcescens*: A potential pathogen for fish. *J. Fish Dis.* 15, 15–26. doi: 10.1111/j.1365-2761.1992.tb00632.x

Bernal, P., Llamas, M. A., and Filloux, A. (2018). Type VI secretion systems in plant-associated bacteria. *Environ. Microbiol.* 20, 1–15. doi: 10.1111/1462-2920.13956

Bertelli, C., and Brinkman, F. S. L. (2018). Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics* 34, 2161–2167. doi: 10.1093/bioinformatics/bty095

Besler, K. R., and Little, E. L. (2017). Diversity of *Serratia marcescens* strains associated with cucurbit yellow vine disease in Georgia. *Plant Dis.* 101, 129–136. doi: 10.1094/PDIS-05-16-0618-RE

Bogard, R. W., Davies, B. W., and Mekalanos, J. J. (2012). MetR-regulated *Vibrio cholerae* metabolism is required for virulence. *MBio* 3:e00236-12. doi: 10.1128/mBio.00236-12

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17:238. doi: 10.1186/s13059-016-1108-8

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., et al. (2005). VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* 33, D325–D328. doi: 10.1093/nar/gki008

Chen, X., Hitchings, M. D., Mendoza, J. E., Balanza, V., Facey, P. D., Dyson, P. J., et al. (2017). Comparative genomics of facultative bacterial symbionts isolated from European *Orius* species reveals an ancestral symbiotic association. *Front. Microbiol.* 8:1969. doi: 10.3389/fmicb.2017.01969

Chung, W.-C., Chen, L.-L., Lo, W.-S., Kuo, P.-A., Tu, J., and Kuo, C.-H. (2013). Complete genome sequence of *Serratia marcescens* WW4. *Genome Announc.* 1:e0012613. doi: 10.1128/genomeA.00126-13

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Czub, M. P., Zhang, B., Chiarelli, M. P., Majorek, K. A., Joe, L., Porebski, P. J., et al. (2018). A Gcn5-related N-acetyltransferase (GNAT) capable of acetylating polymyxin B and colistin antibiotics *in vitro*. *Biochemistry* 57, 7011–7020. doi: 10.1021/acs.biochem.8b00946

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

Demidenko, E. (2018). The next-generation K-means algorithm. *Stat. Anal. Data Min.* 11, 153–166. doi: 10.1002/sam.11379

Dessì, A., Puddu, M., Testa, M., Marcialis, M. A., Pintus, M. C., and Fanos, V. (2009). *Serratia marcescens* infections and outbreaks in neonatal intensive care units. *J. Chemother.* 21, 493–499. doi: 10.1179/joc.2009.21.5.493

Ding, W., Baumdicker, F., and Neher, R. A. (2018). panX: Pan-genome analysis and exploration. *Nucleic Acids Res.* 46:e5. doi: 10.1093/nar/gkx977

Dupriez, F., Rejasse, A., Rios, A., Lefebvre, T., and Nielsen-LeRoux, C. (2022). Impact and persistence of *Serratia marcescens* in *Tenebrio molitor* larvae and feed under optimal and stressed mass rearing conditions. *Insects* 13:458. doi: 10.3390/insects13050458

Egorova, D. A., Solovyev, A. I., Polyakov, N. B., Danilova, K. V., Scherbakova, A. A., Kravtsov, I. N., et al. (2022). Biofilm matrix proteome of clinical strain of *P. aeruginosa* isolated from bronchoalveolar lavage of patient in intensive care unit. *Microb. Pathog.* 170:105714. doi: 10.1016/j.micpath.2022.105714

Escribano, E., Saralegui, C., Moles, L., Montes, M. T., Alba, C., Alarcón, T., et al. (2019). Influence of a *Serratia marcescens* outbreak on the gut microbiota establishment process in low-weight preterm neonates. *PLoS One* 14:e0216581. doi: 10.1371/journal.pone.0216581

Fodor, A., Palkovics, L., and Végh, A. (2022). First report of *Serratia marcescens* from oleander in Hungary. *Phytopathol. Mediterr.* 61, 311–317. doi: 10.36253/phyto-13354

Fortier, L.-C., and Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* 4, 354–365. doi: 10.4161/viru.24498

Friman, M. J., Eklund, M. H., Pitkälä, A. H., Rajala-Schultz, P. J., and Rantala, M. H. J. (2019). Description of two *Serratia marcescens* associated mastitis outbreaks in Finnish dairy farms and a review of literature. *Acta Vet. Scand.* 61:54. doi: 10.1186/s13028-019-0488-7

Gambino, A. S., Déraspe, M., Álvarez, V. E., Quiroga, M. P., Corbeil, J., Roy, P. H., et al. (2021). *Serratia marcescens* SCH909 as reservoir and source of genetic elements related to wide dissemination of antimicrobial resistance mechanisms. *FEMS Microbiol. Lett.* 368:fnab086. doi: 10.1093/femsle/fnab086

Garmory, H. S., and Titball, R. W. (2004). ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. *Infect. Immun.* 72, 6757–6763. doi: 10.1128/IAI.72.12.6757-6763.2004

Gillis, A., Rodríguez, M., and Santana, M. A. (2014). *Serratia marcescens* associated with bell pepper (*Capsicum annuum* L.) soft-rot disease under greenhouse conditions. *Eur. J. Plant Pathol.* 138, 1–8. doi: 10.1007/s10658-013-0300-x

Gosselin, S., Fullmer, M. S., Feng, Y., and Gogarten, J. P. (2022). Improving phylogenies based on average nucleotide identity, incorporating saturation correction and nonparametric bootstrap support. *Syst. Biol.* 71, 396–409. doi: 10.1093/sysbio/syab060

Gust, A. A. (2015). Peptidoglycan perception in plants. *PLoS Pathog.* 11:e1005275. doi: 10.1371/journal.ppat.1005275

Haddy, R. I., Mann, B. L., Nadkarni, D. D., Cruz, R. F., Elshoff, D. J., Buendia, F. C., et al. (1996). Nosocomial infection in the community hospital: Severe infection due to *Serratia* species. *J. Fam. Pract.* 42, 273–277.

Hasan, M. F., Islam, M. A., and Sikdar, B. (2020). First report of *Serratia marcescens* associated with black rot of *Citrus sinensis* fruit, and evaluation of its biological control measures in Bangladesh. *F1000Res.* 9:1371. doi: 10.12688/f1000research.27657.2

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

Izaguirre-Mayoral, M. L., Lazarovits, G., and Baral, B. (2018). Ureide metabolism in plant-associated bacteria: Purine plant-bacteria interactive scenarios under nitrogen deficiency. *Plant Soil* 428, 1–34. doi: 10.1007/s11104-018-3674-x

Jaeger, T., and Mayer, C. (2008). The transcriptional factors MurR and catabolite activator protein regulate N-acetylmuramic acid catabolism in *Escherichia coli*. *J. Bacteriol.* 190, 6598–6608. doi: 10.1128/JB.00642-08

Kamata, R., Matsumoto, K., Okamura, R., Yamamoto, T., and Maeda, H. (1985). The serratial 56K protease as a major pathogenic factor in serratial keratitis. Clinical and experimental study. *Ophthalmology* 92, 1452–1459. doi: 10.1016/s0161-6420(85)33855-1

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

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

Kędzior, M., Seredyński, R., and Gutowicz, J. (2016). Microbial inhibitors of cysteine proteases. *Med. Microbiol. Immunol.* 205, 275–296. doi: 10.1007/s00430-016-0454-1

Khan, A. R., Park, G.-S., Asaf, S., Hong, S.-J., Jung, B. K., and Shin, J.-H. (2017). Complete genome analysis of *Serratia marcescens* RSC-14: A plant growth-promoting bacterium that alleviates cadmium stress in host plants. *PLoS One* 12:e0171534. doi: 10.1371/journal.pone.0171534

Kitts, P. A., Church, D. M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., et al. (2016). Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44, D73–D80. doi: 10.1093/nar/gkv1226

Kotoky, R., and Pandey, P. (2020). Rhizosphere assisted biodegradation of benzo(a)pyrene by cadmium resistant plant-probiotic *Serratia marcescens* S2I7, and its genomic traits. *Sci. Rep.* 10:5279. doi: 10.1038/s41598-020-62285-4

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

Kurz, C. L., Chauvet, S., Andrès, E., Aurouze, M., Vallet, I., Michel, G. P. F., et al. (2003). Virulence factors of the human opportunistic pathogen *Serratia marcescens* identified by in vivo screening. *EMBO J.* 22, 1451–1460. doi: 10.1093/emboj/cdg159

Lancaster, L. J. (1962). Role of *Serratia* species in urinary tract infections. *Arch. Intern. Med.* 109, 536–539. doi: 10.1001/archinte.1962.03620170034005

Létoffé, S., Ghigo, J. M., and Wandersman, C. (1994). Iron acquisition from heme and hemoglobin by a *Serratia marcescens* extracellular protein. *Proc. Natl. Acad. Sci. U.S.A.* 91, 9876–9880. doi: 10.1073/pnas.91.21.9876

Li, P., Kwok, A. H. Y., Jiang, J., Ran, T., Xu, D., Wang, W., et al. (2015). Comparative genome analyses of *Serratia marcescens* FS14 reveals its high antagonistic potential. *PLoS One* 10:e0123061. doi: 10.1371/journal.pone.0123061

Liu, L., Tan, S., Jun, W., Smith, A., Meng, J., and Bhagwat, A. A. (2009). Osmoregulated periplasmic glucans are needed for competitive growth and biofilm formation by *Salmonella enterica* serovar Typhimurium in leafy-green vegetable wash waters and colonization in mice. *FEMS Microbiol. Lett.* 292, 13–20. doi: 10.1111/j.1574-6968.2008.01462.x

Liu, W., Kang, R., Lim, K. L., and Tan, E. K. (2020). Complete genome sequence of *Serratia marcescens* FY, isolated from *Drosophila melanogaster*. *Microbiol. Resour. Announc.* 9:e00755-20. doi: 10.1128/MRA.00755-20

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* 1079, 155–170. doi: 10.1007/978-1-62703-646-7_10

Mahlen, S. D. (2011). *Serratia infections*: From military experiments to current practice. *Clin. Microbiol. Rev.* 24, 755–791. doi: 10.1128/CMR.00017-11

Mataseje, L. F., Boyd, D. A., Delport, J., Hoang, L., Imperial, M., Lefebvre, B., et al. (2014). *Serratia marcescens* harbouring SME-type class A carbapenemases in Canada and the presence of blaSME on a novel genomic island, SmarGI1-1. *J. Antimicrob. Chemother.* 69, 1825–1829. doi: 10.1093/jac/dku040

Matteoli, F. P., Passarelli-Araujo, H., Reis, R. J. A., da Rocha, L. O., de Souza, E. M., Aravind, L., et al. (2018). Genome sequencing and assessment of plant growth-promoting properties of a *Serratia marcescens* strain isolated from vermicompost. *BMC Genomics* 19:750. doi: 10.1186/s12864-018-5130-y

Matteoli, F. P., Pedrosa-Silva, F., Dutra-Silva, L., and Giachini, A. J. (2021). The global population structure and beta-lactamase repertoire of the opportunistic pathogen *Serratia marcescens*. *Genomics* 113, 3523–3532. doi: 10.1016/j.ygeno.2021.08.009

Merlino, C. P. (1924). Bartolomeo Bizio's Letter to the most eminent priest, Angelo Bellani, concerning the phenomenon of the red colored polenta. *J. Bacteriol.* 9, 527–543. doi: 10.1128/jb.9.6.527-543.1924

Mir, A., Rotger, L., and Rosselló, F. (2018). Sound Colless-like balance indices for multifurcating trees. *PLoS One* 13:e0203401. doi: 10.1371/journal.pone.0203401

Moradigaravand, D., Boinett, C. J., Martin, V., Peacock, S. J., and Parkhill, J. (2016). Recent independent emergence of multiple multidrug-resistant *Serratia marcescens* clones within the United Kingdom and Ireland. *Genome Res.* 26, 1101–1109. doi: 10.1101/gr.205245.116

Moreira, C. G., Herrera, C. M., Needham, B. D., Parker, C. T., Libby, S. J., Fang, F. C., et al. (2013). Virulence and stress-related periplasmic protein (VisP) in

bacterial/host associations. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1470–1475. doi: 10.1073/pnas.1215416110

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Ono, T., Taniguchi, I., Nakamura, K., Nagano, D. S., Nishida, R., Gotoh, Y., et al. (2022). Global population structure of the *Serratia marcescens* complex and identification of hospital-adapted lineages in the complex. *Microb. Genomics* 8:000793. doi: 10.1099/mgen.0.000793

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

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. *Microb. Genomics* 2:e000056. doi: 10.1099/mgen.0.000056

Pan, X., Sun, C., Tang, M., You, J., Osire, T., Zhao, Y., et al. (2020). LysR-Type transcriptional regulator MetR controls prodigiosin production, methionine biosynthesis, cell motility, H(2)O(2) tolerance, heat tolerance, and exopolysaccharide synthesis in *Serratia marcescens*. *Appl. Environ. Microbiol.* 86:e02241-19. doi: 10.1128/AEM.02241-19

Payne, S. M. (1993). Iron acquisition in microbial pathogenesis. *Trends Microbiol.* 1, 66–69. doi: 10.1016/0966-842x(93)90036-q

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

Pye, G. W., Jacobson, E. R., Newell, S. M., Scase, T., Heard, D. J., and Dennis, P. M. (1999). *Serratia marcescens* infection in a gopher tortoise, *Gopherus polyphemus*, and use of magnetic resonance imaging in diagnosing systemic disease. *Bull. Assoc. Reptil. Amphib. Vet.* 9, 8–11. doi: 10.5818/1076-3139.9.4.8

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rainey, P. B. (1999). Adaptation of *Pseudomonas fluorescens* to the plant rhizosphere. *Environ. Microbiol.* 1, 243–257. doi: 10.1046/j.1462-2920.1999.00040.x

Redfern, J., Wallace, J., van Belkum, A., Jaillard, M., Whittard, E., Ragupathy, R., et al. (2021). Biofilm associated genotypes of multiple antibiotic resistant *Pseudomonas aeruginosa*. *BMC Genomics* 22:572. doi: 10.1186/s12864-021-07818-5

Rocha, E. P. C., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18, 291–294. doi: 10.1016/S0168-9525(02)02690-2

Rodriguez, G. M., and Smith, I. (2006). Identification of an ABC transporter required for iron acquisition and virulence in *Mycobacterium tuberculosis*. *J. Bacteriol.* 188, 424–430. doi: 10.1128/JB.188.2.424-430.2006

Sahni, R. D., Amalanathan, R., Devanga Ragupathi, N. K., Mathai, J., Veeraraghavan, B., and Biswas, I. (2016). Complete genome sequence of *Serratia marcescens* U36365, a green pigment-producing strain isolated from a patient with urinary tract infection. *Genome Announc.* 4:e00837-16. doi: 10.1128/genomeA.00837-16

Saidenberg, A. B. S., Teixeira, R. H. F., Astolfi-Ferreira, C. S., Knöbl, T., and Ferreira, A. J. P. (2007). *Serratia marcescens* infection in a swallow-tailed hummingbird. *J. Wildl. Dis.* 43, 107–110. doi: 10.7589/0090-3558-43.1.107

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. S. (2014). tqDist: A library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* 30, 2079–2080. doi: 10.1093/bioinformatics/btu157

Saralegui, C., Ponce-Alonso, M., Pérez-Viso, B., Moles Alegre, L., Escribano, E., Lázaro-Perona, F., et al. (2020). Genomics of *Serratia marcescens* isolates causing outbreaks in the same pediatric unit 47 years apart: Position in an updated phylogeny of the species. *Front. Microbiol.* 11:451. doi: 10.3389/fmicb.2020.00451

Schappe, T., Ritchie, D. F., and Thiessen, L. D. (2019). First report of *Serratia marcescens* causing a leaf spot disease on industrial hemp (*Cannabis sativa*). *Plant Dis.* 104:1248. doi: 10.1094/PDIS-04-19-0782-PDN

Shikov, A. E., Malovichko, Y. V., Nizhnikov, A. A., and Antonets, K. S. (2022). Current methods for recombination detection in bacteria. *Int. J. Mol. Sci.* 23:6257. doi: 10.3390/ijms23116257

Shimuta, K., Ohnishi, M., Iyoda, S., Gotoh, N., Koizumi, N., and Watanabe, H. (2009). The hemolytic and cytolytic activities of *Serratia marcescens* phospholipase A (PhlA) depend on lysophospholipid production by PhlA. *BMC Microbiol.* 9:261. doi: 10.1186/1471-2180-9-261

Sikorowski, P. P., Lawrence, A. M., and Inglis, G. D. (2001). Effects of *Serratia marcescens* on rearing of the tobacco budworm (Lepidoptera: *Noctuidae*). *Am. Entomol.* 47, 51–60. doi: 10.1093/ae/47.1.51

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Snipen, L., and Liland, K. H. (2015). micropan: An R-package for microbial pan-genomics. *BMC Bioinformatics* 16:79. doi: 10.1186/s12859-015-0517-0

Stam, E. (2002). Does imbalance in phylogenies reflect only bias? *Evolution* 56, 1292–1295. doi: 10.1111/j.0014-3820.2002.tb01440.x

Stancu, M. M. (2016). Response mechanisms in *Serratia marcescens* IBB(Po15) during organic solvents exposure. *Curr. Microbiol.* 73, 755–765. doi: 10.1007/s00284-016-1108-7

Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N., et al. (2020). Phigaro: High-throughput prophage sequence annotation. *Bioinformatics* 36, 3882–3884. doi: 10.1093/bioinformatics/btaa250

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988

Suryawanshi, R. K., Patil, C. D., Borase, H. P., Narkhede, C. P., Salunke, B. K., and Patil, S. V. (2015). Mosquito larvicidal and pupaecidal potential of prodigiosin from *Serratia marcescens* and understanding its mechanism of action. *Pestic. Biochem. Physiol.* 123, 49–55. doi: 10.1016/j.pestbp.2015.01.018

Tang, Y., Chen, H., Lin, Z., Zhang, L., Upadhyay, A., Liao, C., et al. (2022). Evolutionary genomics analysis reveals gene expansion and functional diversity of arylalkylamine N-acetyltransferases in the *Culicinae* subfamily of mosquitoes. *Insect Sci.* 30, 569–581. doi: 10.1111/1744-7917.13100

Tang, Y., Horikoshi, M., and Li, W. (2016). ggfortify: Unified interface to visualize statistical results of popular R packages. *R J.* 8, 478–489. doi: 10.32614/RJ-2016-060

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

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

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

Tripura, C., Sashidhar, B., and Podile, A. R. (2007). Ethyl methanesulfonate mutagenesis–enhanced mineral phosphate solubilization by groundnut-associated *Serratia marcescens* GPS-5. *Curr. Microbiol.* 54, 79–84. doi: 10.1007/s00284-005-0334-1

Vaikuntapu, P. R., Rambabu, S., Madhuprakash, J., and Podile, A. R. (2016). A new chitinase-D from a plant growth promoting *Serratia marcescens* GPS5 for enzymatic conversion of chitin. *Bioresour. Technol.* 220, 200–207. doi: 10.1016/j.biortech.2016.08.055

Wang, Y., Zhang, K., Shi, X., Wang, C., Wang, F., Fan, J., et al. (2016). Critical role of bacterial isochorismatase in the autophagic process induced by *Acinetobacter baumannii* in mammalian cells. *FASEB J.* 30, 3563–3577. doi: 10.1096/fj.201500019R

Whelan, F. J., Hall, R. J., and McInerney, J. O. (2021). Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Mol. Biol. Evol.* 38, 3697–3708. doi: 10.1093/molbev/msab139

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. doi: 10.1007/978-3-319-24277-4

Williams, D. J., Grimont, P. A. D., Cazares, A., Grimont, F., Ageron, E., Pettigrew, K. A., et al. (2022). The genus *Serratia* revisited by genomics. *Nat. Commun.* 13:5195. doi: 10.1038/s41467-022-32929-2

Wu, S., Blackburn, M. B., Mizell, R. F., Duncan, L. W., Toews, M. D., Sparks, M. E., et al. (2021). Novel associations in antibiosis stemming from an insect pupal cell. *J. Invertebr. Pathol.* 184:107655. doi: 10.1016/j.jip.2021.107655

Xie, Z., and Tang, H. (2017). ISEScan: Automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* 33, 3340–3347. doi: 10.1093/bioinformatics/btx433

Yeung, A. T. Y., Torfs, E. C. W., Jamshidi, F., Bains, M., Wiegand, I., Hancock, R. E. W., et al. (2009). Swarming of *Pseudomonas aeruginosa* is controlled by a broad spectrum of transcriptional regulators, including MetR. *J. Bacteriol.* 191, 5592–5602. doi: 10.1128/JB.00157-09

Zhang, S., Chaluvadi, S. R., and Bennetzen, J. L. (2020). Draft genome sequence of a *Serratia marcescens* strain isolated from the pitcher fluids of a *Sarracenia* pitcher plant. *Microbiol. Resour. Announc.* 9:e01216-19. doi: 10.1128/MRA.01216-19

Zhou, Z., Charlesworth, J., and Achtman, M. (2020). Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* 30, 1667–1679. doi: 10.1101/gr.260828.120

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 September 2023
DOI 10.3389/fmicb.2023.1238580

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Arnab Banerjee,
Serampore College, India
Sunil Banskar,
University of Arizona, United States

*CORRESPONDENCE
Juan P. Cárdenas
✉ juan.cardenas@umayor.cl

RECEIVED 12 June 2023
ACCEPTED 21 August 2023
PUBLISHED 14 September 2023

CITATION
González D, Morales-Olavarria M, Vidal-Veuthey B and Cárdenas JP (2023) Insights into early evolutionary adaptations of the *Akkermansia* genus to the vertebrate gut. *Front. Microbiol.* 14:1238580. doi: 10.3389/fmicb.2023.1238580

COPYRIGHT
© 2023 González, Morales-Olavarria, Vidal-Veuthey and Cárdenas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Insights into early evolutionary adaptations of the *Akkermansia* genus to the vertebrate gut

Dámariz González[1], Mauricio Morales-Olavarria[1], Boris Vidal-Veuthey[1] and Juan P. Cárdenas[1,2]*

[1]Centro de Genómica y Bioinformática, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Santiago, Chile, [2]Escuela de Biotecnología, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Santiago, Chile

Akkermansia, a relevant mucin degrader from the vertebrate gut microbiota, is a member of the deeply branched Verrucomicrobiota, as well as the only known member of this phylum to be described as inhabitants of the gut. Only a few *Akkermansia* species have been officially described so far, although there is genomic evidence addressing the existence of more species-level variants for this genus. This niche specialization makes *Akkermansia* an interesting model for studying the evolution of microorganisms to their adaptation to the gastrointestinal tract environment, including which kind of functions were gained when the *Akkermansia* genus originated or how the evolutionary pressure functions over those genes. In order to gain more insight into *Akkermansia* adaptations to the gastrointestinal tract niche, we performed a phylogenomic analysis of 367 high-quality *Akkermansia* isolates and metagenome-assembled genomes, in addition to other members of Verrucomicrobiota. This work was focused on three aspects: the definition of *Akkermansia* genomic species clusters and the calculation and functional characterization of the pangenome for the most represented species; the evolutionary relationship between *Akkermansia* and their closest relatives from Verrucomicrobiota, defining the gene families which were gained or lost during the emergence of the last *Akkermansia* common ancestor (*LAkkCA*) and; the evaluation of the evolutionary pressure metrics for each relevant gene family of main *Akkermansia* species. This analysis found 25 *Akkermansia* genomic species clusters distributed in two main clades, divergent from their non-*Akkermansia* relatives. Pangenome analyses suggest that *Akkermansia* species have open pangenomes, and the gene gain/loss model indicates that genes associated with mucin degradation (both glycoside hydrolases and peptidases), (micro)aerobic metabolism, surface interaction, and adhesion were part of *LAkkCA*. Specifically, mucin degradation is a very ancestral innovation involved in the origin of *Akkermansia*. Horizontal gene transfer detection suggests that *Akkermansia* could receive genes mostly from unknown sources or from other Gram-negative gut bacteria. Evolutionary metrics suggest that *Akkemansia* species evolved differently, and even some conserved genes suffered different evolutionary pressures among clades. These results suggest a complex evolutionary landscape of the genus and indicate that mucin degradation could be an essential feature in *Akkermansia* evolution as a symbiotic species.

KEYWORDS

***Akkermansia*, pangenome, phylogenomics, gene gain/loss model, dN/dS, Tajima D value, mucin degradation**

1. Introduction

Akkermansia is a genus of anaerobic mucin-degrading Gram-negative bacteria from the Verrucomicrobiota phylum (Derrien et al., 2004). *Akkermansia muciniphila*, the first characterized member of this genus, was isolated from human stool with a medium containing mucin as the primary carbon and nitrogen source (Derrien et al., 2004). The second species described in the genus was *A. glycaniphila*, isolated from reticulated python feces (Ouwerkerk et al., 2016a), and from human stool (Lv et al., 2020). More recently, species such as *A. biwaensis*, *A. massiliensis*, *Candidatus* Akkermansia intestinavium, *Ca.* Akkermansia intestinigallinarum and *Ca.* Akkermansia timonensis, have been also proposed (Ndongo et al., 2022). According to these findings, other analyses from multiple metagenomic datasets suggest that there are numerous species related to *A. muciniphila*, and some of them coexist in different samples from human microbiomes (van Passel et al., 2011; Guo et al., 2017; Xing et al., 2019; Lv et al., 2020; Karcher et al., 2021). In addition, *Akkermansia*-like microorganisms have been detected in samples from other different vertebrates, including primates [e.g., lemurs, gorillas (Ley et al., 2008)] or mice (Presley et al., 2010), several orders of mammals (Geerlings et al., 2021), as well as chickens (Belzer and de Vos, 2012) or reptiles (Costello et al., 2010; Ouwerkerk et al., 2016a).

In the human gastrointestinal (GI) tract microbiome, *A. muciniphila* represents 1–4% of the microbial composition, detectable in a considerable fraction of reference populations (Belzer and de Vos, 2012). Due to its mucin-degrading role, this microbe plays a positive role in maintaining a healthy mucous layer and gut barrier integrity (Derrien et al., 2017). In addition, *A. muciniphila* levels were observed to decrease in inflammatory bowel disease (Png et al., 2010). Furthermore, *Akkermansia* abundance also decreases in subjects with metabolic disorders such as obesity and diabetes, as seen in studies with both murine models and human cohorts [reviewed in Derrien et al. (2017)]. Moreover, intervention studies in mice (Everard et al., 2013) and proof-of-concept studies in human subjects (Depommier et al., 2019) showed that *A. muciniphila* intake could improve parameters associated with insulin resistance, such as glycemia or insulin sensibility, among other factors. These findings showed the potential of *A. muciniphila* as a new probiotic agent and a source of postbiotics (Vinderola et al., 2022).

From an evolutionary perspective, *Akkermansia* is an interesting case of a member of the GI microbiota with close relatives inhabiting non-related environments. Furthermore, *Akkermansia*-related species are almost the only members of the Verrucomicrobiota phylum identified in the vertebrate gut microbiome, especially in mammals (Youngblut et al., 2020; Levin et al., 2021). The members of the Verrucomicrobiota (formerly Verrucomicrobia) phylum (Hedlund, 2015) are Gram-negatives with additional features, such as the presence of intracellular compartments bounded by internal membranes. The Verrucomicrobiota comprises microorganisms from various ecological niches, including sponge symbionts and soil inhabitants (Kamneva et al., 2012). The main taxonomic classes include three formally defined groups: Verrucomicrobiae, Spartobacteria, and Opitutae, in addition to two other putative groups, including formally undefined microorganisms. According to *Bergey's Manual, Akkermansia* is the only characterized genus of the Akkermansiaceae family, a member of the Verrucomicrobiales order (Hedlund and Derrien, 2015); other genera found in the Verrucomicrobiales are *Verrucomicrobium, Prosthecobacter, Luteolibacter, Roseibacillus, Percisirhabdus*, and *Rubritalea*. All those groups mentioned above comprise free-living bacteria from soil and marine environments, except for *Rubritalea*, a genus comprising sponge symbionts (Hedlund et al., 2015). Both 16S rRNA-based phylogeny, and phylogenomic approaches confirmed the phylogenetic relationship of *Akkermansia* with other members from Verrucomicrobiales (Kamneva et al., 2012).

The phylogenetic context of *Akkermansia* in the Verrucomicrobiota taxonomic tree can be utilized as a case study of how ancient genomic changes sculpted the origin of a gut colonizer. Previous studies have proposed that habitat changes, rather than the phylogenetic background, are preponderant factors defining the pangenome content of a microbial group (Maistrenko et al., 2020). This feature may implicate that *Akkermansia* experienced massive evolutionary changes during its emergence and separation from their non-gut colonizer relatives, including massive gene gain and loss events, or also the emergence of different evolutionary pressures on certain lineage-specific genes. However, it is yet to be established how early or late those events occurred to make *Akkermansia* a professional gut colonizer. The implications of the elucidation of these transformations can be harnessed not only for evolutionary reasons but also for biotechnology implications, such as the discovery of genetic factors that could improve gut colonization or the confirmation of metabolic abilities involved in the proper gut colonization process.

In this study, in order to understand the (pan)genomic transformations involved in the divergence of a gut colonizer from free-living relatives, we performed a phylogenomic analysis of a set of more than three hundred *Akkermansia* genomes and metagenome-assembled genomes (MAGs), calculating the pangenome of different *Akkermansia* genomic species, searching for genes that could be involved in the differentiation of this genus, and calculating the evolutionary pressure of several genes from the most represented genomic species.

2. Methodology

2.1. Genome dataset definition

A set of six hundred candidate genomes classified the Akkermansiaceae family was downloaded from the NCBI Genbank FTP site (May 2022). Those genomes were evaluated using two criteria: their taxonomic identity and their degree of completeness and contamination. The taxonomic identity was confirmed by using the program '*classify_wf*' of the GTDB-TK program, version 2.1.0 (Chaumeil et al., 2019), using the database release 207 as the reference, selecting all genomes classified into the *Akkermansia* genus ("g__ Akkermansia"). Genome completeness and contamination were calculated using the program '*lineage_wf*' from CheckM version 1.1.3 (Parks et al., 2015); only those genomes with completeness equal to or higher than 90%, and contamination below 5%, were selected, according to current recommendations (Bowers et al., 2017).

2.2. Annotation, identification of orthogroups, and phylogenomic tree

All members of the final dataset were annotated *de novo* using Prokka, version 1.11 (Seemann, 2014) (relevant parameters:

--metagenome --kingdom Bacteria --addgenes). Orthogroups from the set of the final dataset were calculated by Orthofinder version 2.5.5 (Emms and Kelly, 2019) with the '*-og*' parameter. In order to make a phylogenetic tree for the *Akkermansia* plus the outgroup genomes, a multiple sequence alignment was constructed from a set of 87 concatenated, single-copy conserved orthogroups by using MAFFT version 7.490 (parameters: *--maxiterate 1000 --localpair*) (Katoh et al., 2019); the alignment was used by *iqtree* version 2.1.4 (Nguyen et al., 2015) (parameters: *-m TEST --alrt 1000*) to generate a maximum likelihood-based tree with an aLRT with 1,000 replicates as the branch support test. The phylogenomic tree was visualized using the *Toytree* Python package (Eaton, 2020).

2.3. Definition of genomic species

In order to detect the genomic species represented in the *Akkermansia* selected dataset, we combined the prediction from GTDB-tk (see above) with the prediction of clusters defined by average nucleotide identity (ANI) values. All genomes were compared in an all-vs-all manner using FastANI version 1.32 (Jain et al., 2018) with default parameters. The raw pairwise comparison data was filtered, discarding all ANI values below 95%, the classical intra-species boundary for microbial genomes (Richter and Rosselló-Móra, 2009). Filtered pairwise comparisons were analyzed by the *MCL* program, creating putative genomic species clusters in an analogous manner as observed in network clustering (van Dongen and Abreu-Goodger, 2012).

2.4. Pangenome analysis

In order to analyze the pangenome of the main *Akkermansia* genomic species groups, separated Orthofinder executions were performed with the proteomes from each cluster without any outgroup (parameters: *-M msa -y*). The orthogroup matrix (including unassigned orthogroups) was obtained for each run and utilized for different pangenome metrics. Pangenome curves were created using the *panplots* function in R (created by *SioStef*)[1], using 1,000 permutations. The γ value from the Power law mentioned by Tettelin et al. (2008) was also calculated for each pangenome, using the function *curve_fit* from *scipy* python package, using the equation "$y = ax^{\gamma}$," applied on the *panplots* output. We calculated the shell, cloud, "soft-core" and core components of the pangenome from the complete orthogroup matrix by using Python scripts with the *pandas* package, considering the following criteria: *core* gene families as the orthogroups present in 100% of the strains, soft-core groups as present in between 99.999 and 90% of the strains, shell as groups present between 89 and 15% of strains, and cloud as the gene families present in between 14% and the equivalent to two strains. Unique groups can be deduced from the set of "species-specific orthogroups," and the "unassigned genes," both reported by Orthofinder. Figures were created with *ggplot2* and the *ggarrangment* packages.

2.5. Functional annotation

Akkermansia proteomes were analyzed using EggNOG mapper version 2.1.6 (Cantalapiedra et al., 2021) (parameters: "--tax_scope_mode narrowest --tax_scope prokaryota_broad --go_evidence experimental"). COGs at the *root* level were extracted for those classifications, and the categories were established according to the current COG version.[2] Carbohydrate active enzymes (Flint et al., 2012), including glycoside hydrolases (GH), were searched using the HMM database (v. 11) from dbCAN, the search tool based on CAZy (Zhang et al., 2018), using an e-value <1e-10. Proteases were searched in the dataset using the MEROPS Database Release 12.4 (Rawlings et al., 2018). Sulfatases involved in mucin desulfation were searched by using the SulfAtlas database version 2.3.1 as the reference (Stam et al., 2023). MEROPS and SulfAtlas searches were performed by using Diamond (Buchfink et al., 2015) *blastp* searches using the *Akkermansia* protein dataset as the query and the full-length sequence repository in each case as the database (e-value <1e-10, subject coverage >70%).

2.6. Gene gain/loss model for combined pangenomes

In order to create a general gene gain/loss model for *Akkermansia*, the most conserved (core + soft-core) of each species cluster (or the current proteome for 1-genome species groups) were compared by using Orthofinder as previously mentioned. The generated orthogroup matrix was utilized for the generation of a phylogenomic tree, as previously mentioned, confirming that the position of the species in the tree was the same as the one observed in the previous tree. This tree and the binary version of the complete orthogroup matrix were used by the software Count (Csurös, 2010) for the calculation of gene gain/loss rates following the Csûrös - Miklós model, optimized with a Poisson distribution at the root; the rates were also optimized considering a variation across families to 1:1:1:1 gamma categories for the edge length, the loss rate, gain rate, and the duplication rate, respectively. The convergence criteria were set to a likelihood delta of 0.05 with a maximum of 1,000 rounds. The calculated rates were used to generate an analysis following Wagner parsimony using the same penalty score (equal to 1) for gains and losses. The final tree was represented by using the ETE3 python package (Huerta-Cepas et al., 2016). By using ETE3 as well, the tree structure and the Count output were compared in order to identify the orthogroups present and absent in each LCA or leaf (extant genomic species). The orthogroups were also compared to the EGGNOG-mapper output to define the functional content of each LCA.

2.7. Evolutionary metrics: Tajima D and dN/dS ratios

In order to calculate Tajima's D value, the nucleotide sequences from the coding sequences of those proteins found in the core and

1 https://github.com/SioStef/panplots/

2 https://www.ncbi.nlm.nih.gov/research/cog

soft-core genes from the top 4 pangenomes analyzed above were utilized. Sequences were aligned with MAFFT, as mentioned above. Nucleotide multiple alignments were utilized to calculate Tajima's D values using the *tajima.test* function from *pegas* R package.[3] Graphs were created with *ggplot2*.

In order to calculate dN/dS ratios, pairwise values were calculated for an aligned set of orthogroups conserved among four genomic species (against the CDS content from *A. glycaniphila* PyT as the outgroup). Alignments were converted into codon-aware aligned using PAL2NAL (Suyama et al., 2006). Calculation of dN, dS, and ω was made by using CODEML program from the PAML package (Yang, 2007), using the following parameters: "*runmode = −2, seqtype = 1, CodonFreq = 3, model = 1, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, omega = 0.5*." Additionally, pairwise Tajima-Nei distances were calculated from the same members of each alignment using the *Bio::Align::DNAStatistics* module from BioPerl (Stajich, 2007). Pairwise comparisons with distances equal to zero, dN/dS > 5, and dS > 10 were discarded from the analysis. Pairwise dN/dS and Tajima-Nei distance values were also filtered, considering only orthogroup comparisons between sequences from different genomic species.

2.8. Horizontal gene transfer (HGT) calculation

Genes potentially acquired by horizontal gene transfer (HGT) were inferred using HGTector v2.0b3 (Zhu et al., 2014). Selected coding sequences were analyzed against the default reference database (38,488 genomes, retrieved in September 2022); the database was formatted using DIAMOND (Buchfink et al., 2015), retrieving valid hits using default parameters. Sequence hits from Akkermansiaceae (NCBI TaxID 1647988) were considered as *self*.

2.9. Statistical tests

All statistical tests were conducted in R version 4.1.0. Normal distribution was evaluated using the Shapiro–Wilk test (*shapiro.test*). We compared data groups from two conditions with non-normal distribution using the Mann–Whitney *U* test (*wilcox.test*).

3. Results

3.1. The phylogenetic composition of the filtered *Akkermansia* dataset

The final dataset, obtained from the NCBI Genbank repository and filtered by taxonomy and completeness/contamination standards, corresponds to a collection of 367 *Akkermansia* genomes distributed in 25 genomic species clusters (Supplementary Figure 1 and Supplementary Tables 1, 2). Some interesting genomes, such as *Ca.* Akkermansia intestinigallinarum (GCA_019114365.1), could not be included in the final dataset due to their below-cutoff completeness (84.61%, against the UID2982 marker set). Most species clusters were previously detected in the GTDB-TK reference. The four most prevalent clusters represented *Akkermansia muciniphila*, *Akkermansia* sp004167605, *Akkermansia muciniphila*_B, and *Akkermansia* sp001580195. Each cluster contained 193, 73, 40, and 19 genomes, respectively (Supplementary Table 2). In contrast, 13 genomic species groups are represented by only one genome per cluster.

A representative tree was created from the original dataset (Figure 1), including five outgroups representing different genera of Verrucomicrobiales (also listed in Supplementary Table 1). This rooted tree showed the existence of two main clades (Clades A and B, see Figure 1). The clade A comprised *A. muciniphila* and other related species (GTDB names sp900545155, sp905200945, muciniphila_A, muciniphila_B, muciniphila_E, muciniphila_C, sp001580195, sp004167605, and Cluster25, without no specific species level classification in GTDB). Compared with previous studies (Karcher et al., 2021; Lv et al., 2022), the GTDB *A. muciniphila* genomic species corresponds to the *Amuc I* phylogroup, whereas *Akkermansia muciniphila_B* corresponds to *Amuc II*, *Akkermansia muciniphila_C* to *Amuc III*, and *Akkermansia* sp001580195 to *Amuc IV*. On the other hand, Clade B includes *A. glycaniphila, Ca.* Akkermansia intestinavium, *Akkermansia muciniphila_D*, and several other species, including an undetected cluster in GTDB, called Cluster08. Genetic distances between members of Clade A are significantly lower than the observed distances between members of Clade B (non-normal distribution; $p < 0.05$ according to Mann–Whitney *U* test, data not shown). Unless specified, we used this 367-genome dataset (and its corresponding genome species clusters) for all downstream analyses.

3.2. Pangenome properties of the *Akkermansia muciniphila*, sp004167605, *muciniphila*_B, and sp001580195 species clusters

Recently, comparative genomics and *A. muciniphila* pangenome studies have been performed multiple times (van Passel et al., 2011; Guo et al., 2017; Xing et al., 2019; Lv et al., 2020; Geerlings et al., 2021; Karcher et al., 2021; Kim et al., 2022; Li et al., 2022; Lv et al., 2022; Ouwerkerk et al., 2022). However, the use of an Orthofinder-based strategy is not common. In order to assess our strategy in the *Akkermansia* dataset, we performed a pangenome analysis using the four genomic species clusters with the highest number of genomes. For those genomic species, the core-, soft-core, shell, cloud, and unique components and the accumulation curves were calculated from the Orthofinder analysis results, and the γ value from the Power law was also calculated for each pangenome (see above).

This analysis showed that the core genome for those genomes ranged between 2,170 and 1,265 orthogroups (Figure 2 and Table 1). The core genome set from Cluster 2 (*Akkermansia* sp004167605) is the longest among the analyzed species, and Cluster 1 (*A. muciniphila*) contained the smallest core genome. The pangenome comprised between 6,865 and 3,515 orthogroups among those four species; these changes in pangenome content seem to depend on the number of genomes in each cluster, a feature seen in other open pangenome models (Costa et al., 2020). The curve profile of the pangenome accumulation plot (Supplementary Figures 2A–D and Table 1)

3 https://cran.r-project.org/web/packages/pegas/index.html

10.3389/fmicb.2023.1238580

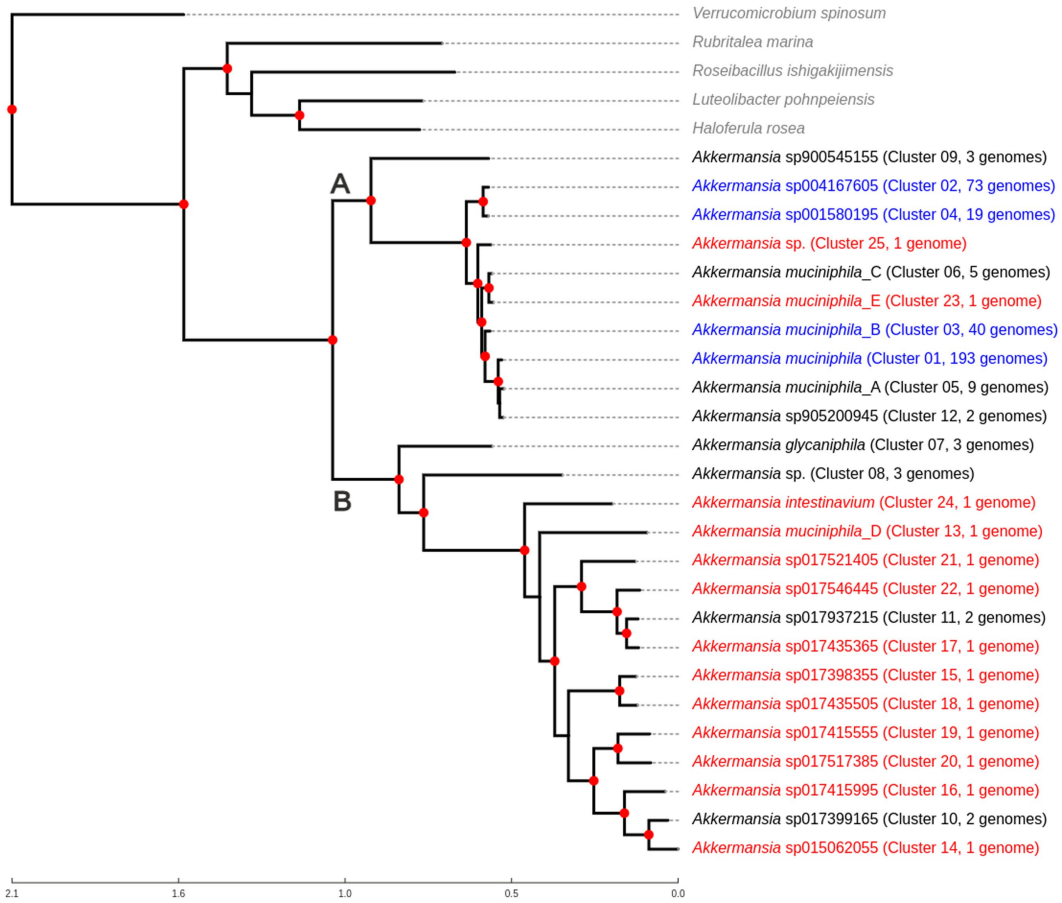


FIGURE 1
Phylogenomic tree showing the relationship between different *Akkermansia* genomic species, in comparison with other related members of Verrucomicrobiales. The tree was created from the alignment of 87 single-copy conserved protein families, using the maximum likelihood method in IQTREE, with the use of the approximate likelihood ratio test (aLRT) as the branch support test. Red nodes indicate branch support values equal to 100. Taxa in red represent genomic species composed only of one genome; taxa in blue indicate that those genomic species are the top four clusters with the most genomes. Clades "A" and "B" are mentioned in the text.

exhibited the properties of an open pangenome ($0 < \gamma < 1$) for all four genomic species, confirming this feature observed in other studies, observing the pangenomes for the Amuc I to Amuc IV phylotypes (Bukhari et al., 2022).

The openness of those four pangenomes requires addressing whether the core/soft-core contents are functionally different compared to the shell, cloud, or unique genes. To observe these differences, we compared the percentage of genes assigned to each COG category between the core genome pangenome and the core, soft-core, cloud, shell, or unique fractions (Supplementary Figures 3A–D), as well to obtain the Log2 ratios between the core and the "accessory" (the sum of soft-core, shell, cloud, and unique) pangenome fractions (Figures 2A–D). Functional categories where the percentage of genes assigned in the core pangenome is two-fold (log2 equal or higher than 1) were considered core-enriched categories. On the other hand, functional COG categories where the percentage of genes assigned to the core pangenome was one-half (log2 equal or lower than −1) of the accessory fraction were considered core-depleted.

These analyses showed some general tendencies; for example, categories associated with metabolism (e.g., categories C, G, E, F, H) and information transfer (e.g., J, L, K) were the most prevalent in the core and soft-core fractions. On the other hand, categories such as X ("Mobilome") or Q ("Secondary metabolite transport and metabolism") had the highest percentages detected in the shell cloud or unique gene fractions. Genes without COGs have strongly higher percentages in the fraction of unique genes (data not shown). Cluster 2 and Cluster 4 were the genomic species with the highest number of core-enriched categories compared to the accessory genes. COG categories U, G, Q, P, C, D, N, F, H, and E showed high core-to-accessory ratios in Cluster 2, while Cluster 4 showed I, K, C, E, Q, F, H, P, J, and N presenting the largest ratios. The categories that overlap these two Clusters correspond to P, C, F, H, and N. Cluster 1 showed that category J was core-enriched, and Cluster 3 showed that P and E categories were core-enriched (Figures 2A–D). Category N ("Cell motility") is highly core-enriched in both Cluster 2 and Cluster 4. Motility-associated genes in those species were related to pili synthesis as an outer-membrane protein, suggesting that these clusters could differentiate from the others due to their differences in host-microorganism membrane interactions (Ottman et al., 2016), which can be specific depending on the host organism. Furthermore, category C was also enriched

10.3389/fmicb.2023.1238580

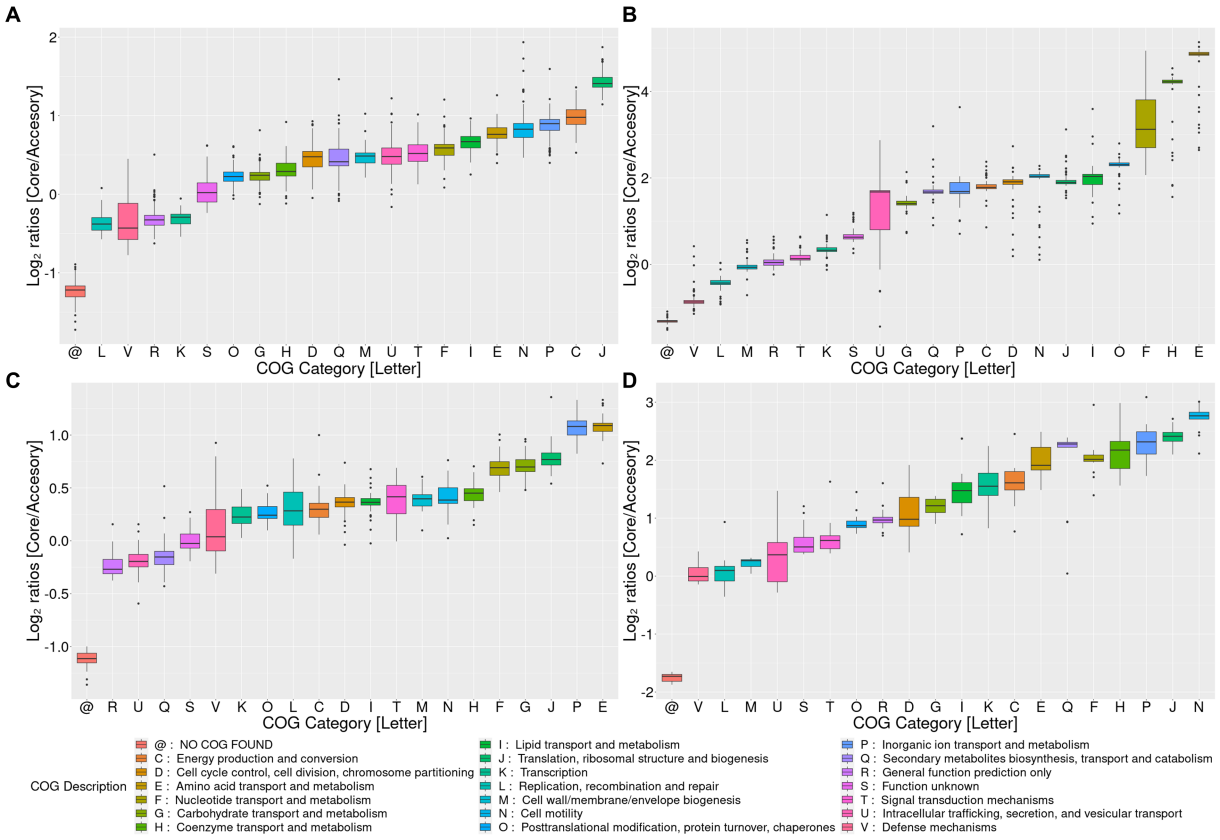


FIGURE 2
Pangenome analysis for the top four *Akkermansia* genomic species clusters with the most members. Log2 ratios of the percentage of core versus accessory (the sum of soft-core, shell, cloud, and uniques) genes for each COG category among pangenomes. Clusters I, II, III, and IV are represented by panels **(A–D)**, respectively. COG category descriptions are represented in the lower panel @: gene families without any COG.

TABLE 1 Main statistics from pangenome analysis for the top four genomic clusters found in the *Akkermansia* dataset.

	Cluster 1 (*A. muciniphila*)	Cluster 2 (A. sp004167605)	Cluster 3 (*A. muciniphila*_B)	Cluster 4 (A. sp001580195)
Total genomes	193	73	40	19
Total pangenome genes	6,865	4,622	4,341	3,515
Core genes (in 100% of genomes)	1,265	2,170	1,462	1,936
Soft-core genes (less than 100%, more than 90%)	666	370	733	249
Shell genes (less than 90%, more than 15%)	642	110	806	769
Cloud genes (less than 15%, at least in two genomes)	1,905	589	551	154
Unique genes (in just one genome)	2,387	1,383	789	407
Gamma value (Power law)	0.2637	0.18199	0.15516	0.10665

in the core group of genes in Clusters 2 and 4, suggesting that "Energy production and conversion" might be one of the key functions in the genomic differentiation of *Akkermansia* species. Since all those four Clusters are strongly related, and Clusters 2 and 4 shared a direct common ancestor (belonging to Lineage A, see Figure 1), those differences could respond to evolutionary differences between those organisms and the other two analyzed clusters. This final observation also suggests the need to analyze the evolutionary changes of the gene content between different organisms across *Akkermansia* evolution.

10.3389/fmicb.2023.1238580

3.3. Gene gain/loss model analysis of *Akkermansia* lineages and the reconstruction of the gene composition and functionality of *LAkkCA* (the Last *Akkermansia* Common Ancestor)

As seen previously, the 367 *Akkermansia* genome dataset utilized in this study represented 25 potential genomic species (Figure 1), as some of their pangenome properties could be related to evolutionary patterns (see above). To study the evolutive gene gain/loss dynamics across the evolution of this genus, we predict the content of the different common ancestors, from the most recent common ancestors to the putative "Last *Akkermansia* Common Ancestor" (from now, *LAkkCA*), in comparison with a set of members of *Rubritalea*, *Roseibacillus*, *Luteolibacter* and *Haloferula*, the closest Verrucomicrobiales as outgroups (Figures 1, 3 and Supplementary Figure 4). This comparison was performed with core and soft-core gene sets within species groups since this comparison focused on ancestral changes rather than more recent changes.

According to the prediction, *LAkkCA* contained 1,419 genes, including 169 gained genes compared to its non-*Akkermansia* ancestor (Figure 3). The prediction also showed that the last common ancestors (LCAs) from Lineages A and B (Figure 1) experienced subsequent gene gain/loss events: 176 gained/12 missed genes in Lineage-A LCA, and 52 gained/67 missed genes in Lineage-B LCA. The median of gene loss events in Lineage A was lower than the median in Lineage B ($p < 0.05$ according to Mann–Whitney *U* test, data not shown). This feature suggests that Lineage A was more conservative than Lineage B in its gene content. In general, gene acquirements were most frequent than gene loss events across the *Akkermansia* phylogeny.

The functional profile of those genes showed that the gene content of *LAkkCA* (Figure 4A), expectedly, included an important percentage of genes in several essential processes such as Translation (category J), DNA Replication (category L), and Transcription (category K), as a signal of the role of those essential processes in cellular configuration. Genes involved in metabolism (COG categories C, E, F, G, H, I) were found in important proportions. Notoriously, genes involved in Carbohydrate Metabolism (Category G, Figure 4B) corresponded to a significant fraction of genes acquired in *LAkkCA*; in counterpart, genes from the J, K, and L categories were detected in lower fractions among the acquired genes (Figure 4B). This latter observation concurs with previous analyses in different models showing that those kinds of functions were most reluctant to be transferred horizontally, at least between distant groups (Kanhere and Vingron, 2009). Additionally, a gene set of 142 genes, without any COG assignment, was detected in *LAkkCA*.

In addition to the gene set acquired by *LAkkCA*, there is a set of genes that *LAkkCA* lost during its differentiation from their non-Akkermansia relatives (Figure 4C). According to the gene gain/loss model, 496 genes were lost by the ancestor of the *Akkermansia* genus (Figure 3 and Supplementary Figure 4). These changes could be associated with adaptions to the new niche; for example, whereas nearly %1 of the genes acquired by *LAkkCA* belonged to the "Transcription" (K) category (Figure 4B), almost 10% of the missed genes by *LAkkCA* belonged to this same category (Figure 4C), suggesting that the change of niche could involve changes in the transcriptional regulation program of the emergent genus.

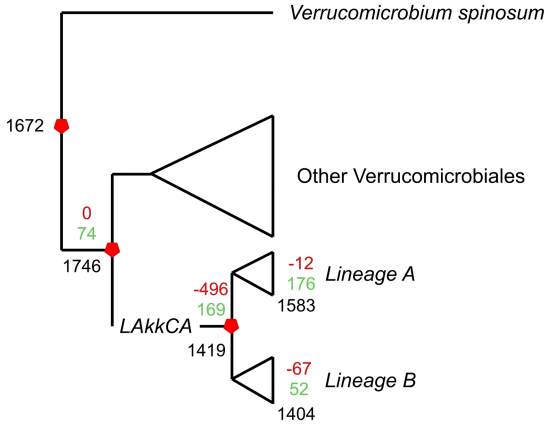


FIGURE 3
Overview of the gene gain/loss model representing gene content changes across the evolution of the *Akkermansia* genus. The phylogenomic tree from Figure 1 was combined with the orthogroup matrix generated from the Orthofinder output, and analyzed by Count. Black numbers represent the number of shared families for a given ancestor or extant genome; green and red numbers represent gained or missed gene families, respectively.

3.4. Predicted metabolism and genetic features of *LAkkCA*

The prediction of the putative genetic content that harbored the hypothetical last common ancestor of this complete genus raises the opportunity to predict the properties of early members of the *Akkermansia* genus. We annotate the putative gene content from *LAkkCA* with current state-of-the-art tools to proceed with this prediction. The predicted functional repertoire of *LAkkCA* (Supplementary Table 3) includes the capability to encode genes for the NADH:quinone dehydrogenase, the Cytochrome bd-type quinol oxidase [a terminal oxidase with high oxygen affinity (Borisov et al., 2021)], and the ATPase complexes, in addition to a relatively complete central carbon metabolism (including glycolysis, a partial TCA cycle, and the reductive branch of the pentose phosphate pathway). The genetic capabilities of *LAkkCA* also include the biosynthesis of some phospholipids, such as phosphatidyl glycerol, phosphatidylserine, phosphatidylethanolamine, and cardiolipin. The prediction of *LAkkCA* metabolism also suggests the production of acetate via pyruvate dehydrogenase [EC:1.2.5.1] (K00156) and the production of propionyl-CoA via propionyl CoA:succinate CoA transferase (COG0427). The possession of NADH dehydrogenase, ATPase, and some terminal oxidases suggest an inheritance of the respiratory metabolism from their non-*Akkermansia* ancestors, something noticeable if we consider that previous studies suggested that *A. muciniphila* may conduct microaerobic metabolism in the mucus layer niche (Ouwerkerk et al., 2016b), even if is also capable of performing anaerobic metabolism. In concordance with this ability to use oxygen, *LAkkCA* also contained several genes involved in oxidative stress: several genes encoding members of COG0526 (Thiol-disulfide isomerase or thioredoxin), in addition to a member of COG1225 (Bcp Peroxiredoxin), COG0450 (AhpC peroxiredoxin), COG1592 (Rubreythrin), COG0605 (Superoxide dismutase) and COG0753 (KatE Catalase). This latter finding suggests the need for a proper adaptation to higher oxygen levels during the origin of the *Akkermansia* genus.

10.3389/fmicb.2023.1238580

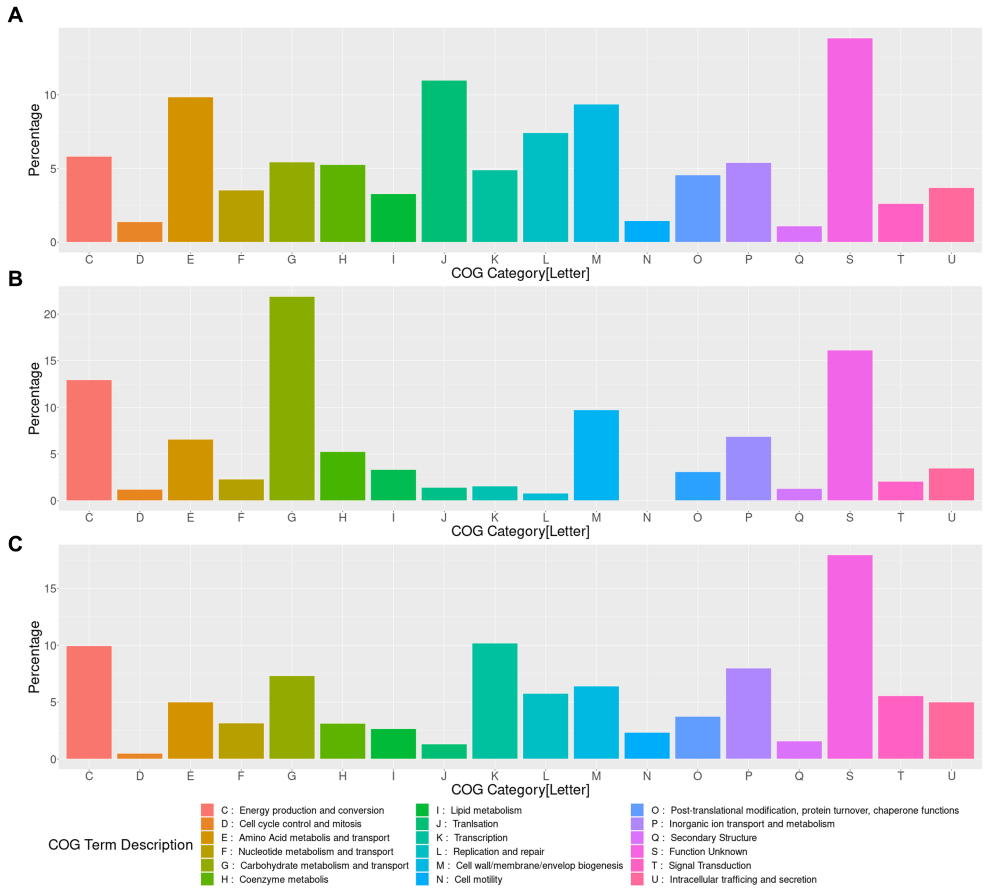


FIGURE 4
Functional characterization of the genes content present, gained and lost in *LAkkCA*. **(A)** Percentage of assigned genes to each COG category in the predicted total gene content of *LAkkCA*. **(B)** Percentage of assigned genes to each COG category in the sub-dataset of gained genes in *LAkkCA* (genes present in this LCA but absent from its ancestral node). **(C)** Percentage of assigned genes to each COG category in the sub-dataset of those genes absent in *LAkkCA*, in comparison with their ancestor node. COG category descriptions are represented in the lower panel.

According to the prediction of its genetic content, *LAkkCA* also contained a series of proteins involved in its relationship with the host and environment. For example, this ancestor contained a set of genes encoding proteins containing COG0666 ("Ankyrin repeat," ANKR), a ubiquitous domain involved in protein–protein interactions. In Bacteria, ANKRs were mostly investigated in proteobacterial organisms, especially pathogens, where those proteins could be associated with protein secretion systems involved in pathogenic interactions with the host (Al-Khodor et al., 2010). In the case of *Akkermansia*, those proteins could be part of a set of secreted proteins that could generate an effect on the host, as seen with other secreted and exposed proteins detected in *A. muciniphila* (Vidal-Veuthey et al., 2022). *LAkkCA* also encoded a predicted set of genes for Type IV pilus assembly, suggesting a role of this ubiquitous complex, involved in several functions such as motility, biofilm formation, and adherence (Ligthart et al., 2020), in the early adaptation of this genus to the gut environment. Moreover, *LAkkCA* also encodes a couple of genes encoding putative autotransporters; those proteins contain their own exportation system, transporting one domain (the "passenger" domain) of the same protein across the outer membrane of Gram-negative bacteria (Clarke et al., 2022). Our prediction suggests that those two autotransporters may be involved in mucin degradation and adherence, since their passenger domains were a sialidase (GH33) domain and an adhesin, respectively (data not shown).

3.5. Insights into the mucin metabolism in *LAkkCA* and its evolution across the *Akkermansia* genus

As mentioned earlier, mucin metabolism is a distinctive feature of *Akkermansia* species, and some studies have suggested that *A. muciniphila* showed higher growth in mucin than in glucose (Derrien et al., 2004; Glover et al., 2022). A set of carbohydrate-active enzymes (CAZymes) are commonly associated with glycan moieties degradation in mucin. Those CAZymes are important players in the gut microbiome, and several key GHs are identified in *Akkermansia* involved in its ability to degrade mucin (Chen et al., 2019; Glover et al., 2022). We searched for enzymes involved in the degradation of this glycoprotein in *LAkkCA*, the predicted ancestors for Lineages A and B, and the representative content for each species cluster (Supplementary Table 4). We found that genes encoding some CAZyme families involved in mucin degradation, such as GH20 (related to β-hexosaminidases and β-1,6-N-acetylglucosaminidases),

GH29 (related to α-fucosidases), GH33 (related to neuraminidase and sialidases), GH35 (β-galactosidases/β-glucosaminidases), and GH95 (also related to α-L-fucosidases), were detected in *LAkkCA* (Figure 5). Moreover, those same functions were remarkably absent from the non-*Akkermansia* relatives (data not shown). In this ancestor, we also found another CAZyme family called GH16_3. This GH family comprises O-glycanase capable of targeting the polyLacNAc structures (consists of repeated N-acetyllactosamine units) within oligosaccharide side chains of both animal and human mucins (Crouch et al., 2020). Most GHs found in *LAkkCA* were also found most, if not all, *Akkermansia* species, experienced in some cases the inclusion of multiple copies, such as the case of GH20 (Supplementary Table 4). These analyses strongly suggest that the origin of *Akkermansia* involved the possession of a basic set of glycan mucin degradation, and that gene set was strongly conserved or even amplified.

Subsequently, despite being absent in *LAkkCA*, other relevant CAZymes such as GH2 (mainly composed by beta-galactosidases), GH84 (N-acetyl β-glucosaminidase), GH89 (α-N-acetylglucosaminidase), and GH110 (α-galactosidase), were also found in *Akkermansia* species as expected (Glover et al., 2022). For example, a copy for GH2 was predicted in Lineage-A LCA and among different *Akkermansia* species, GH2 was also detected, often in multiple copies (Supplementary Table 4). The amplification of the number of copies of certain GHs involved in mucin may reflect the reinforcing role of these enzymes in this important feature. Moreover, it may also reflect clade-specific or lineage-specific adaptations from different *Akkermansia* species to adapt to the mucin complexity found in different animals, from reptiles to mammals (Belzer and de Vos, 2012; Ottman et al., 2017). The research of phylogenetic patterns in those GH families may gain some insights into the processes of functional diversification that *LAkkCA* suffered until the current day representatives.

CAZymes are not the only relevant component to search in the mucin degradation machinery. Since mucin is a glycoprotein, the peptidase repertoire is also important to explore among *Akkermansia* genomes. Additionally, the ability of the gut microbiome to degrade peptides is a well-reported phenomenon (Wallace and McKain, 1997). Protease activities in the human gut are relevant for several microbial community members since they could be involved in the use of nutrients, host interactions, and even connections with host health and disease (Caminero et al., 2023). To execute this search, the same representative genetic content from the 25 genomic species, as well as the content of *LAkkCA* and the Lineages A- and B- LCAs, were compared against MEROPS (Rawlings et al., 2018). Since mucin is commonly sulfated, we also searched for sulfatases by using SulfAtlas as a reference (Stam et al., 2023). The search for peptidases and sulfatases associated with mucin metabolism in *Akkermansia* species showed that several of those enzymes were conserved among the genus and found in *LAkkCA* (Supplementary Table 5). For example, peptidases from the metallopeptidase family (M03A, M15D, M20F, M24A, M24B, M38, M42, M50B), serine proteases (S26A, S33, S41A, S54), cysteine proteases (C26, C82A) and other types (U32, I04, A24A, T05), and were detected in *LAkkCA*. Some of them, such as S26A (signal peptidase I), A24A (type 4 prepilin peptidase), and S41A (C-terminal processing peptidase-1), for example, seem to have housekeeping functions. In contrast, some families such as U32 (collagenase), I04 (alpha-1-peptidase inhibitor), or M42 (related to

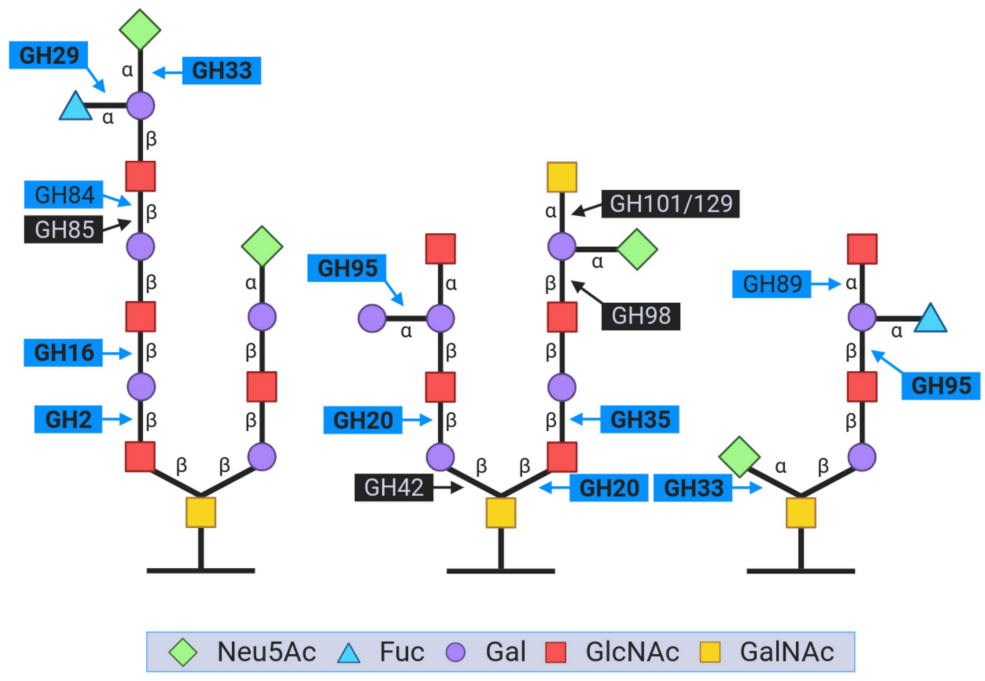


FIGURE 5
Representation of the key glycoside hydrolase (GH) families involved in the breakdown of carbohydrate moieties during mucin degradation. The figure is inspired by the content from Glover et al. (2022). The GH names in bold are those that were found in *LAkkCA*. GH names closed by blue boxes corresponded to the enzymes detected in *Akkermansia* genomes. GH names closed by black boxes corresponded to the enzymes absent in *LAkkCA* and *Akkermansia* genomes.

tetrahedral aminopeptidases), seem to be acquired as special adaptations. For example, bacterial collagenases have been associated with pathogenic bacteria (Duarte et al., 2016), but no information is available about the role of this peptidase in gut commensals. However, the U32 family was also detected in members of *Prevotella* and *Paraprevotella*, common gut inhabitants, suggesting a common role in the gut environment (Patra and Yu, 2022). Notably, peptidases from the M03A family were found not only in *LAkkCA* and in Lineages A and B LCAs, but also present in multiple copies in all the 25 species, in an analogous manner like GH20. This feature may suggest that this particular peptidase family could have a role in the ability of *Akkermansia* species to colonize the gut, or also that these peptidases could be involved in mucin degradation. The functional importance of these proteases in *Akkermansia* physiology could be related to their ability to interact with host cells, although their exact role remains to be established. The search for sulfatases showed that all members, including *LAkkCA*, Lineages A and B LCAs and the core/representative genome from all *Akkermansia* species, contained a uniform, well-conserved sulfatase patrimony (Supplementary Table 5), including members of sulfatase subfamilies S1-4 (containing [Colonic mucin]-endo-D-Galactose-3-sulfate 3-O-sulfohydrolases), S1-11 (including [mucin]-exo-N-acetyl-D-glucosamine-6-sulfate 6-O-sulfohydrolases), S1-15 (including [mucin]-D-Galactose-6-sulfate/N-acetyl-D-galactosamine-6-sulfate 6-O-sulfohydrolase), S1-16 (mucin-D-Galactose-4-sulfate/N-acetyl-D-galactosamine-4-sulfate 4-O-sulfo hydrolase) and S1-20 (inclding mucin- D-Galactose-3-sulfate 3-O-sulfohydrolases and N-acetyl-D-galactosamine-3-sulfate 3-O-sulfatases). This finding has a strong coincidence with the case of the CAZymes: peptidases putatively involved in mucin metabolism have strong conservation among the genus and even are predicted to be present in the last common ancestor of the *Akkermansia* genus.

3.6. Prediction of horizontal gene transfer (HGT) events in the different members of the *Akkermansia* genus

The acquisition of new genes can be explained by three main evolutionary processes, *de novo* gene gain, gene duplication, and HGT (Douglas and Langille, 2019). The birth of *de novo* genes can occur at any time in the evolutionary history of a microorganism, detecting these events in ancient genes is complex since it has been shown that rapidly evolving orthologous genes are often not detected in distant taxa, and they tend to be misclassified as *de novo* genes (Elhaik et al., 2006; Van Oss and Carvunis, 2019). Likewise, it has been shown that the adaptation of bacteria to new environments is mainly mediated by the expansion of protein families encoded by genes obtained through HGT and not by gene duplication (Treangen and Rocha, 2011). In this regard, HGT is a source of phenotypic innovation and an important niche adaptation mechanism (Wolska, 2003; Ravenhall et al., 2015). Genetic conjugation, transduction, and transformation are key processes within HGT (Ravenhall et al., 2015), through which genetic material is exchanged between microorganisms that share the same microenvironment despite not sharing a vertical ancestry (Soucy et al., 2015). Therefore, HGT provides a potential adaptive advantage in the bacterium accepting the genetic material since it allows rapid gene transfer between distantly related species (Douglas and Langille, 2019), and it seems to be a reasonable alternative to investigate genetic traits acquired by a bacterial population. In this context, we performed an analysis using the HGTector tool, which detects possible genes derived from HGT events to identify the probable taxonomic origin of genes with a considerable signal for HGT. In this case, the Akkermansiaceae family level was considered as the "self group" to detect donors only from clearly distinctive taxonomic groups.

We predicted genes potentially from HGT for the 25 clusters representing the 367 annotated genomes of *Akkermansia* species. The analysis showed that nearly 6.9 to 10.3% of the predicted genes in the core/soft-core of the cluster 1 to 12 were predicted with an HGT signal, whereas the percentage of genes putatively received by HGT in clusters 13 to 25 ranged between 6.63 to 8.76%. In all genomic species, putative horizontally transferred genes exhibited similar behavior in their putative donors (Figure 6). For example, in all groups, more than 50% of the genes were predicted to have an unresolvable donor from the *Bacteria* domain or from a "cellular organism" (NCBI TaxID 1), followed by members from the *Proteobacteria* and *Bacteroidota* phyla. Interestingly, we detected in the clusters of *Akkermansia* spp. different genes associated with families of glycosyl hydrolase enzymes (GH16, GH31, GH35, and GH57), with a predicted donor from *Bacteria* (superkingdom) and *Bacteroidota*. Some of these enzymes were found in intestinal bacteria such as *Bacteroides plebeius*, *Bifidobacterium longum* subsp. *infantis* and *Fibrobacter* spp. (Qi et al., 2005; Hehemann et al., 2010; Tarracchini et al., 2021), which accounts for possible HGT events. Likewise, in several clusters, detected genes linked to an efflux pump (AcrAB) were also identified with a potential donor from *Proteobacteria*. This efflux pump confers resistance against a wide variety of antimicrobial compounds, such as bile salts, by expelling them out of the cell (Sun et al., 2014). Additionally, AcrAB has been reported in *Escherichia coli* of intestinal origin (Ma et al., 1995), which indicates the probable HGT between different bacterial genera. Consequently, HGT is a process of adaptation to a habitat shared by taxonomically diverse bacterial populations (Zaneveld, 2011; Chen et al., 2021).

HGTector does not determine a direction for gene acquisition from HGT. This is due to the limitations of the reference databases used (Chen et al., 2021), as well as the lack of a systematic phylogenetic analysis, which would help to strengthen the determination of the potential taxonomic origin of the genes acquired by HGT (Zhu et al., 2014). Another drawback is that this type of analysis does not include information regarding the synteny of groups of genes, which, being present in different bacterial taxa may reflect the action of HGT (Lawrence and Roth, 1996; Imam et al., 2011), information on transposable genetic elements is not included either, which is important, since some transposons are shared by Gram-negative and positive bacteria, and could shed light on a possible diversification in the acquisition of genes through HGT (Ojo et al., 2004). In this scenario, it is difficult to trace the probable taxonomic origin of possible HGT-derived genes. Therefore, some questions remain open, such as whether it is possible that *Akkermansia* species have obtained a repertoire of genes through HGT from members of the intestinal microbiota or if these genes originated in the genus *Akkermansia* and were transferred to other intestinal microorganisms, or also if the genes transferred to *Akkermansia* spp. they underwent a specialization process and were later transferred to other taxa. Verifying these situations is an intriguing issue and still to be elucidated.

10.3389/fmicb.2023.1238580

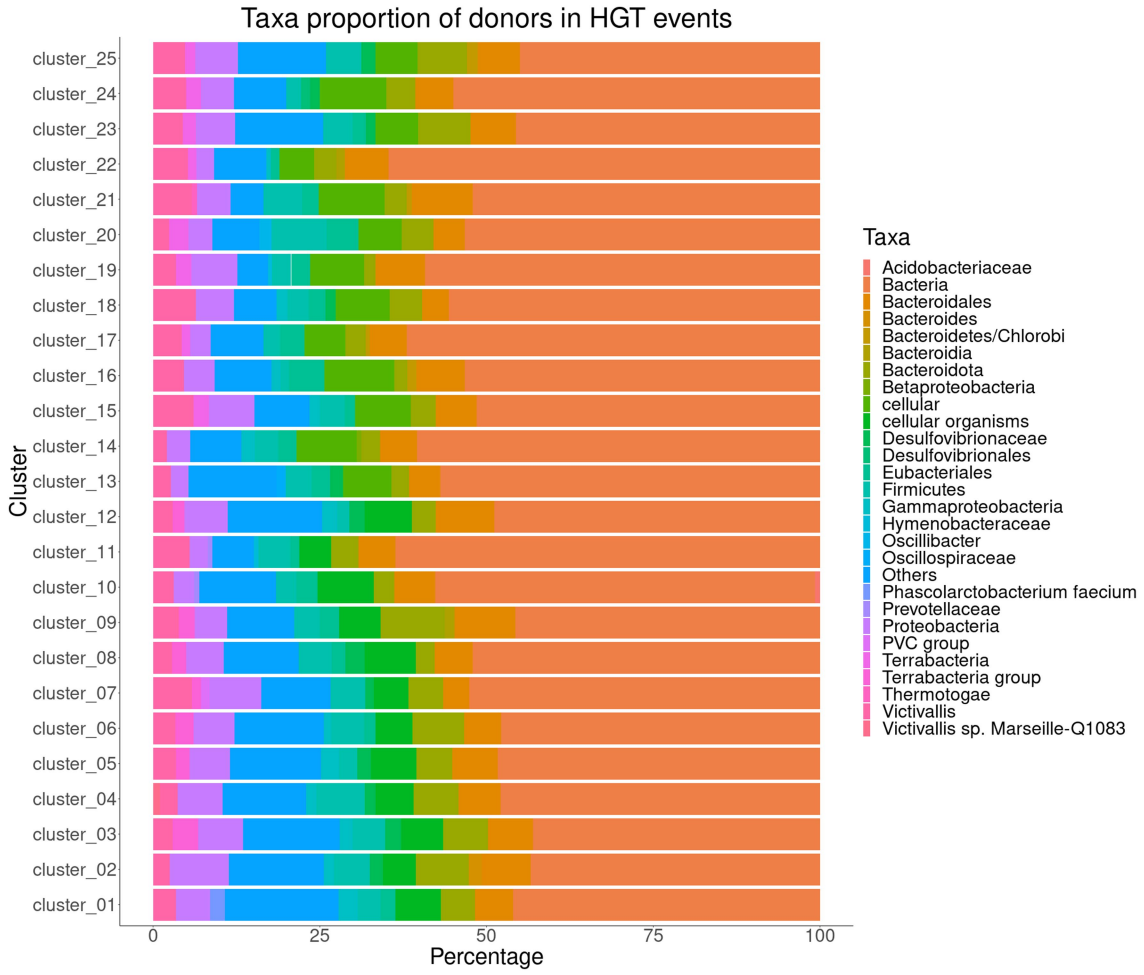


FIGURE 6
Taxa proportion of donors in Horizontal Gene Transfer events predicted among the representative sequences from 25 clusters (involving the core/soft-core from clusters #1 to #12 and the genomes of the only representative from clusters #13 to #25) representing the genomic species found in the *Akkermansia* dataset.

3.7. Tajima distance analyses across the main *Akkermansia* species suggest differential genetic diversity

Tajima D statistical test is a population genetic test used to elucidate if a gene family evolved in a neutral manner, or if it evolved under a non-random process, such as balancing selection, demographic expansion, or contraction, among other effects (Tajima, 1989). In order to evaluate the main tendencies in genetic diversity among the four top *Akkermansia* species clusters (see section 3.2), we calculated the Tajima D values for the core gene families for each species. Our analysis (Figure 7) showed that the distribution of Tajima values had different distributions among the four species (Figure 7A), which Cluster 2 (*Akkermansia* sp004167605) showed a shift to more negative values. Cluster 1 (corresponding to the bonafide *A. muciniphila*) showed a more extended distribution, as seen as well in the boxplots (Figure 7B). In addition, core gene families from Clusters 3 and 4 showed a shift to more positive values in comparison with Cluster 2. The order of median D values was: Cluster1 > Cluster4 > Cluster3 > Cluster2. Using the Mann–Whitney *U* test to make comparisons among two groups, we can found that Clusters #1 and #2 exhibited significant statistical differences between Tajima D values ($p < 2.07\times10^{-156}$), with a common language effect size (CLES) value of 0.809, which means that 80.9% probability that randomly chosen observation from one group will be greater than a randomly selected from the other group, making this difference substantial. In the same way, in comparison between cluster #1 and cluster #3 and between cluster #1 and cluster #4, both comparisons showed significant differences (*p*-values: 9.59×10^{-47} and 6.01×10^{-12}, respectively). However, CLES values show that both comparisons suggest that differences are more modest (CLES of 0.674 and 0.578, respectively). Based on the aforementioned information, our results show that the differences between clusters are sufficient to back up the potential description of new *Akkermansia* species. Differences between cluster #2 versus cluster #3, and cluster #2 versus cluster #4 were also significantly different, although with lower CLES values (*p*-values: 1.35×10^{-81} and 8.12×10^{-246}; CLES: 0.30 and 0.18, respectively). This information suggests that the more negative distribution in Cluster 2 is significant in relation to the other three clusters.

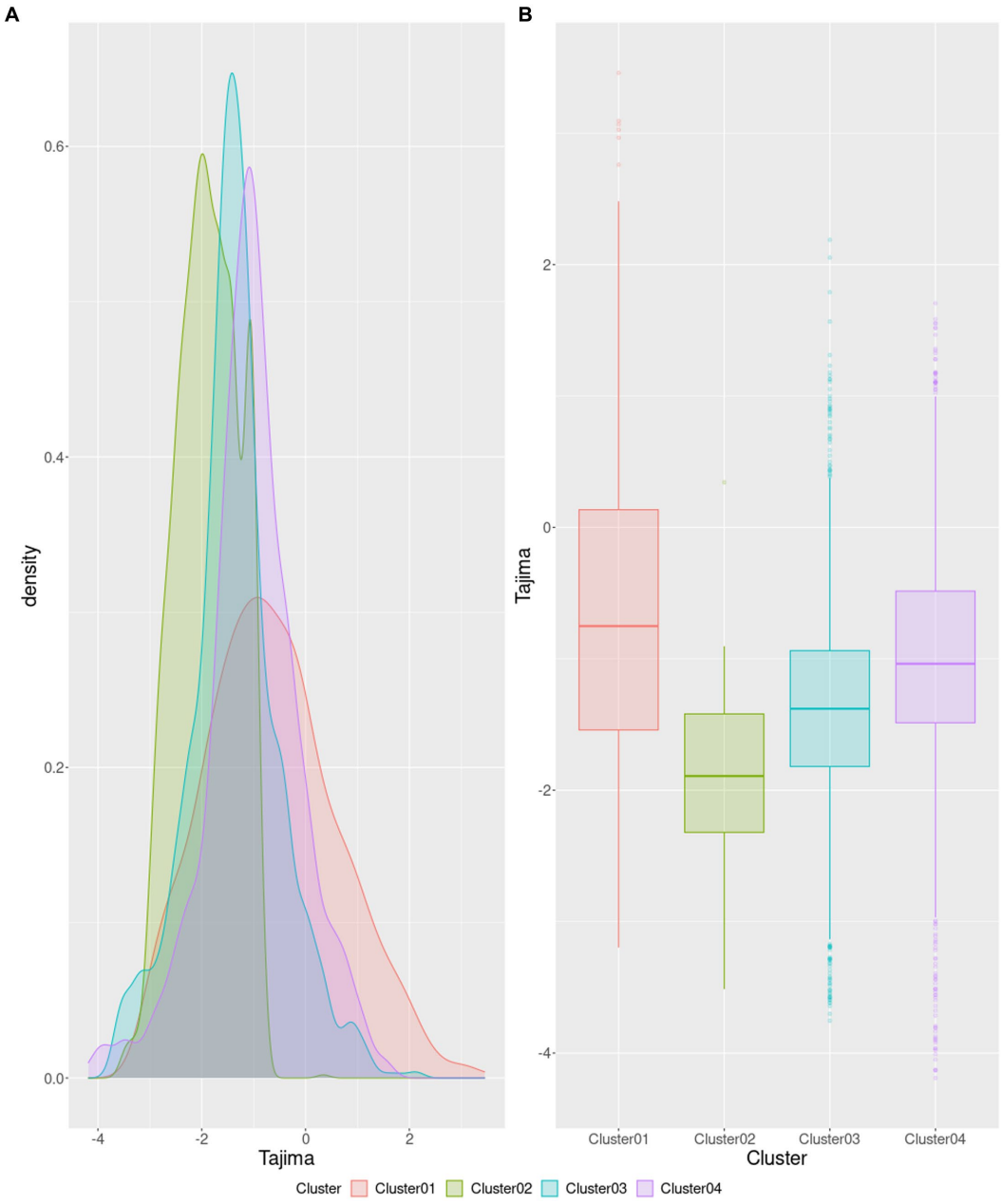


FIGURE 7
Tajima D distribution among the core gene content among different *Akkermansia* genomic species, presented as density curves **(A)** and boxplots **(B)**.

The observed more negative distribution in the core set from members of cluster 2 may reflect the properties of their genetic diversity. Highly negative Tajima values (below −2, with a significant *p*-value) may reflect the effect of positive selection since it indicates an excess of rare alleles, suggesting population expansion. In counterpart, very positive D values (above 2) suggest a high proportion of common alleles, indicating balancing selection or population contraction (Carlson et al., 2005). In the case of Cluster #1, the high amplitude of the distribution could suggest a more neutral evolutionary pattern among gene families, although this may also be affected by sample size (Cluster #1 have the most genomes). Therefore, the available information on those four *Akkermansia* genomic species suggests that each one exhibits different selection pressures, as a reflection of their roles in the GI tract.

Since these gene families (orthogroups) were annotated with EggNOG mapper, a posterior approach was the classification of Tajima value distribution across different COGs categories based on each cluster (Supplementary Figure 5). It is possible to observe that for Clusters #1 and #4, a great proportion of categories showed a distribution toward moderate values (closer to zero), suggesting that DNA-sequence changes were nearly neutral in those species clusters (Carlson et al., 2005). However, in Clusters #2 and #3, this analysis showed that genes related to the defense mechanism in bacteria or cell motility (V and N categories, respectively) were not following the neutral theory of molecular evolution. These two categories are commonly found to unfollow the neutral theory of molecular evolution in bacterial genomes; for example, some genes related to CRISPR are in a constant arms race,

which involves a process of coevolution between bacteriophages and bacteria interaction (Takeuchi et al., 2012). In the same way, genes in a COG category like N (including genes related to transfer events between bacterial cells) are under the effect of purifying selection, which involves a tendency to molecular change at the gene level (N'Guessan et al., 2021).

3.8. Comparing gene selection pressures among four *Akkermansia* species groups found orthogroups with significantly higher selection among species

Different gene diversities among the main *Akkermansia* species represented in the genome dataset raise the possibility that conserved genes among different species groups have different selective pressures. In order to elucidate that aspect, the pairwise dN/dS ratio was compared from the core, single-copy, orthogroups between a symmetrical amount of genomes ($n = 19$) from those mentioned above the top four *Akkermansia* genomic species. Since dN/dS ratios are metric values for macroevolution, all comparisons were contrasted using the genomic content from *A. glycaniphila* as the outgroup; this ensured that all comparisons were compared with each other using representatives from the same outgroup. Additionally, all dN/dS comparison datasets between representative genomes (having a non-normal distribution, confirmed by the Shapiro–Wilk test) were compared using the Mann–Whitney *U* test, discarding all groups without significant differences.

The comparison between the same set of conserved orthogroups, between two different *Akkermansia* species (Figure 8) shows that despite several orthogroups did not have differential pressure signals, a set of few orthogroups have differentially different evolutionary pressures among species groups. Between Clusters #1 and #2, two orthogroups were found to have a remarkably higher dN/dS median in Cluster #2 against Cluster #1, whereas three orthogroups exhibit the inverse behavior (Figure 8A). When clusters #1 and #3 were compared, only one orthogroup was found exceptionally higher in Cluster #3 (Figure 8B); finally, when clusters #1 and #4 were compared, only one orthogroup was found exceptionally higher in Cluster #4 and one another in Cluster #1 (Figure 8C). The list of the functions represented in those remarked orthogroups were involved in a set of functions such as ribosome assembly and function, energy metabolism, and mucin degradation (Table 2).

The composition of a bacterial ribosome is defined as a macromolecule, and due to this different parts could play a specific role in protein synthesis (Lin et al., 2018), in this context, the effects of natural selection could be different depending on the process, either for environmental or antibiotic effects (Byrgazov et al., 2013; Lin et al., 2018). An orthogroup predicted to encode the ribosomal protein L16, was reported with a remarkably positive selection in Cluster #2, in comparison with Cluster #1 (Figure 8A and Table 2). This protein is an important component that helps to stabilize ribosomal structure (Nishimura et al., 2004). Mutations in this protein could be related to the indiscriminate use of antibiotics (Adrian et al., 2000; Gomez et al., 2017), due to the effect these drugs could have on the gut microbiome (Patangia et al., 2022). This study also showed that an orthogroup encoding the NADH-quinone oxidoreductase subunit B (Figures 8B,C and Table 2), which plays a role in energy production and is a multisubunit integral membrane enzyme that participates in different types of respiratory chains (aerobic and anaerobic) and contributes to survival or energy conservation in a variety of lifestyles (Spero et al., 2015). This orthogroup showed a remarkably more positive selection in comparison with clusters #1 versus #3, and #1 versus #4. In this case, this protein could be under positive selection due to the adaptation to environmental changes (Jayaraman et al., 2022).

Interestingly, one of the orthogroups with significantly more positive pressure among species encoded for a beta-hexosaminidase, a protein involved in mucin degradation (Figure 8A and Table 2). As previously mentioned, different members of the gut microbiota interact with the mucus that covers and protects the gastrointestinal epithelium, being capable of degrading glycans that are part of the mucus (Sauvaitre

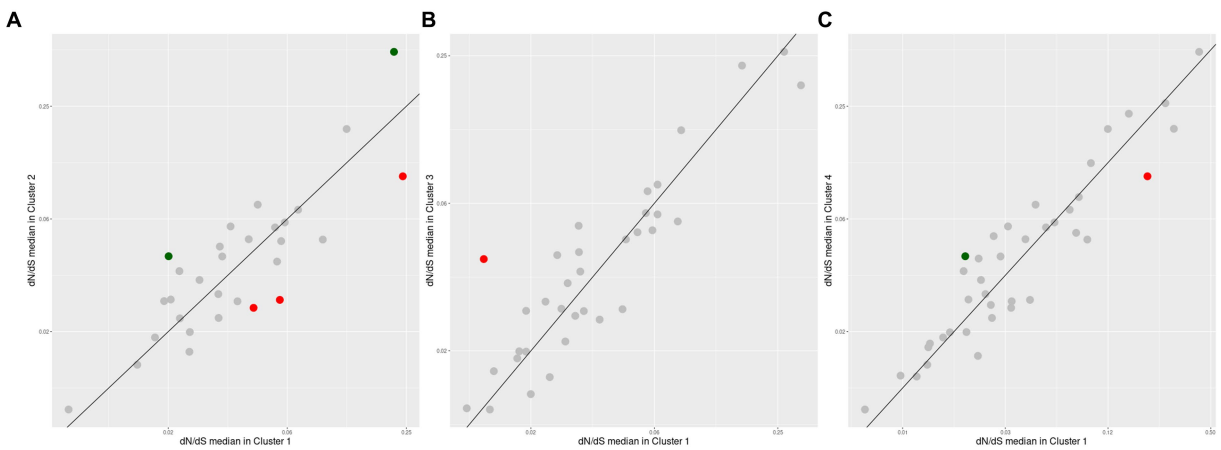


FIGURE 8
Differential positive selection effects in conserved orthogroups among the top four *Akkermansia* genomic species. Each point is an orthogroup found among two different species which values are compared with the same outgroup (*A. glycaniphila*, type strain). The diagonal line indicates the zone when median pairwise dN/dS in both lineages have the same value. Dots with red or green colors represent orthogroups with remarkably different dN/dS medians (abs(Log2 ratio) > 1, $p < 0.05$, Mann-Whithey *U* test). Comparison corresponds to orthologous from Cluster #1 versus Cluster #2 **(A)**, Cluster #1 versus Cluster #3 **(B)**, and Cluster #1 versus #4 **(C)**.

TABLE 2 List of orthogroups found with remarkably different dN/dS medians (abs(Log2 ratio) > 1, $p < 0.05$, Mann-Whithey *U* test) across the top four *Akkermansia* species clusters.

Cluster comparison	Feature	OG	Log2 ratio	Orthogroup function
#1 versus #2	Higher in cluster #1	817	1.33	NADH-quinone oxidoreductase subunit B-1
	Higher in cluster #1	863	1.17	NADH-quinone oxidoreductase subunit B-4
	Higher in cluster #2	475	−1.01	Ribosomal protein L16
	Higher in cluster #2	1,357	−1.18	Beta-hexosaminidase
	Higher in cluster #2	1,298	−1.31	NADH-quinone oxidoreductase subunit B-2
#1 versus #3	Higher in cluster #3	1,365	2.01	NADH-quinone oxidoreductase subunit B-3
#1 versus #4	Higher in cluster #1	817	1.12	NADH-quinone oxidoreductase subunit B-1
	Higher in cluster #4	863	−1.01	NADH-quinone oxidoreductase subunit B-4

et al., 2021). Mucin is a glycoprotein, and a major component of the mucus layer covering the intestinal epithelium (Becker et al., 2022). Hexosaminidases (members of the GH20 family) is an enzyme that catalyzes the hydrolysis of glycosidic linkages, catalyzing the cleavage of terminal β-D-GlcNAc and β-D-GalNAc residues (Xu et al., 2020), which correspond to a glycoside hydrolases 20 (GH20). Genes encoding this protein were found to have a more positive selection in Cluster #2 versus #1. This effect on selection may reflect the special role of this member of the GH20 in the gut microbiome functional network, where different microbes compete to degrade mucin (Kostopoulos et al., 2021), and this more diversifying selection pressure is a consequence of the need for more variability between different *Akkermansia* lineages to adapt their ability to degrade the carbohydrate moieties of mucin in different contexts, or moreover, to respond to a coevolution process between the glycan structural diversity in the host and glycoside hydrolases (Sonnenburg et al., 2005).

4. Discussion

Akkermansia is a very interesting group of gut microbiome inhabitants in various vertebrates. This study demonstrated the existence, at the genomic level, of a set of at least 25 species of this genus, including several previously undescribed variants. The presence of several 1-genome clusters suggests the need for the sequencing and discovery of more *Akkermansia* isolates and MAGs. Our evolutionary analysis also showed that the genetic diversity observed through different relevant species clusters through the Tajima D analysis supports the existence of different evolutionary pressures among the member of the genus. The pangenome analysis confirmed that *Akkermansia* species have an open pangenome, suggesting that more gene content diversity remains to be discovered. On the other hand, dN/dS ratio suggests that genes conserved in different lineages show different patterns of natural selection, which can potentially indicate a degree of specialization of different families of proteins that these genes encode, such as the GH20 family, which is related to mucin degradation, and is diversified in the genus.

The use of a gene gain/loss model could show the properties of a hypothetical entity, called *LAkkCA*, that could be the founder member of the genus. A predicted respiratory metabolism, a complete set of carbon metabolism, and a basic set for mucin degradation suggest that *LAkkCA* could be a mucin degrader with microaerobic metabolism, suggesting that this feature is profoundly rooted to the origin of the genus. Despite several studies comparing *Akkermansia* genomes (as previously mentioned), there are only a few studies focused on the evolutionary dynamics of gene gain/loss events [for example (Kim et al., 2022)]. Moreover, no study has been performed considering the underestimated taxonomic diversity of the genus, and no insight into the origin of the genus has been made. Since *Akkermansia* is a genus detected in the GI tract microbiota from mammals, birds, reptiles, or even some amphibians (Zhang et al., 2020), it is reasonable that *Akkermansia* could start to colonize vertebrate GI tract as near as vertebrates appeared. This hypothesis is compatible with the notion that mucins, the main carbon source for described *Akkermansia* species, is a glycoprotein strongly conserved among vertebrates: MUC2, the primary mucin in the mammalian GIT, contained homologs in all vertebrates (Lang et al., 2016). Moreover, it has been proposed that gel-forming mucins could have originated previously from the origin of vertebrates, being found in other metazoans, such as members of Cnidaria, Porifera, and Ctenophora (Lang et al., 2016). Mucins, as important roleplayers in GIT maintenance (Grondin et al., 2020), are important points of contact with microbial activity and metabolism. It is well known that, as a specialized mucin-degrader, *Akkermansia* species are capable of promoting epithelial development in the intestine (Kim et al., 2021) and eliciting, in some conditions, mucin production itself (Shin et al., 2014), among other effects. In the colon, mucins could form a multi-layered structure, with the outer side involved in the interactions with the microbiome (Johansson et al., 2011). Our study showed that, even within the 25 species clusters of *Akkermansia* (comprising a higher taxonomic diversity than previously shown), there is a very conserved gene set involving in mucin degradation and potential interactions between the bacterium and its host, including potential adhesion pathways or special exported proteins mediating surface interactions (such ankyrin proteins) and extracelulluar matrix degradation (such collagenases). These findings support the hypothesis that *Akkermansia* could born inside the GI tract and the ability to degrade mucin could shape its evolution.

In addition, the fact that most of the genes were cataloged in the *Bacteria* domain (through HGT), opens the way to the possibility that the genus *Akkermansia* obtained genes (e.g., GHs) from various intestinal taxa, or they were created by *Akkermansia* spp. and have been transferred to other microorganisms, or even that genes

transferred to *Akkermansia* have been perfected through contact with various intestinal microenvironments of different hosts, to be later transferred to different taxa.

Our results show the necessity of new studies with the goal to understand the unique new species and the intraspecific evolutionary processes of *Akkermansia*, as well as, the existence of a copious taxonomic diversity for this genus. A systematic phylogenetic analysis must be implemented to decode the probable taxonomic origin of several HGT-derived genes.

Finally, the need to explore and discover more *Akkermansia* isolates from several other members of the vertebrates may help to gain more insight into the relationship between mucin, the evolution of vertebrates, and the natural history of *Akkermansia*. Moreover, the understanding of the evolutionary process involved in the origin of the machinery to process mucin in *Akkermansia* spp. could help to understand the relationship with the maintenance of gut health, as well as their improvement through the design of probiotics.

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 author.

Author contributions

JC conceived and directed the study. DG, MM-O, BV-V, and JC analyzed the data. All authors collaboratively elaborated, edited, corrected the text and figures in the manuscript, read the manuscript, and approved the content.

Funding

JC was supported by ANID Fondecyt Project #11200209. BV-V was supported by *ANID Doctorado Nacional*/2021-21211564.

Acknowledgments

The authors want to thank Jose Luis Maturana for his contributions in the early stage of this project.

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.2023.1238580/full#supplementary-material

References

Adrian, P. V., Zhao, W., Black, T. A., Shaw, K. J., Hare, R. S., and Klugman, K. P. (2000). Mutations in ribosomal protein L16 conferring reduced susceptibility to evernimicin (SCH27899): implications for mechanism of action. *Antimicrob. Agents Chemother.* 44, 732–738. doi: 10.1128/AAC.44.3.732-738.2000

Al-Khodor, S., Price, C. T., Kalia, A., and Abu Kwaik, Y. (2010). Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol.* 18, 132–139. doi: 10.1016/j.tim.2009.11.004

Becker, H. E. F., Kameli, N., Rustichelli, A., Heijnens, B. A. M., Stassen, F., Penders, J., et al. (2022). Higher *in vitro* mucin degradation, but no increased paracellular permeability by faecal water from Crohn's disease patients. *BioRxiv*. doi: 10.1101/2022.08.26.505386

Belzer, C., and de Vos, W. M. (2012). Microbes inside--from diversity to function: the case of Akkermansia. *ISME J.* 6, 1449–1458. doi: 10.1038/ismej.2012.6

Borisov, V. B., Siletsky, S. A., Paiardini, A., Hoogewijs, D., Forte, E., Giuffrè, A., et al. (2021). Bacterial oxidases of the cytochrome bd family: redox enzymes of unique structure, function, and utility as drug targets. *Antioxid. Redox Signal.* 34, 1280–1318. doi: 10.1089/ars.2020.8039

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi: 10.1038/nbt.3893

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

Bukhari, S. A. R., Irfan, M., Ahmad, I., and Chen, L. (2022). Comparative genomics and pan-genome driven prediction of a reduced genome of *Akkermansia muciniphila*. *Microorganisms* 10:1350. doi: 10.3390/microorganisms10071350

Byrgazov, K., Vesper, O., and Moll, I. (2013). Ribosome heterogeneity: another level of complexity in bacterial translation regulation. *Curr. Opin. Microbiol.* 16, 133–139. doi: 10.1016/j.mib.2013.01.009

Caminero, A., Guzman, M., Libertucci, J., and Lomax, A. E. (2023). The emerging roles of bacterial proteases in intestinal diseases. *Gut Microbes* 15:2181922. doi: 10.1080/19490976.2023.2181922

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

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., et al. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15, 1553–1565. doi: 10.1101/gr.4326505

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Chen, M.-Y., Teng, W.-K., Zhao, L., Hu, C.-X., Zhou, Y.-K., Han, B.-P., et al. (2021). Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *ISME J.* 15, 211–227. doi: 10.1038/s41396-020-00775-z

Chen, X., Wang, J., Liu, M., Yang, W., Wang, Y., Tang, R., et al. (2019). Crystallographic evidence for substrate-assisted catalysis of β-N-acetylhexosaminidas from *Akkermansia muciniphila*. *Biochem. Biophys. Res. Commun.* 511, 833–839. doi: 10.1016/j.bbrc.2019.02.074

Clarke, K. R., Hor, L., Pilapitiya, A., Luirink, J., Paxman, J. J., and Heras, B. (2022). Phylogenetic classification and functional review of autotransporters. *Front. Immunol.* 13:921272. doi: 10.3389/fimmu.2022.921272

Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., and Baraúna, R. A. (2020). First steps in the analysis of prokaryotic pan-genomes. *Bioinform. Biol. Insights* 14:1177932220938064. doi: 10.1177/1177932220938064

Costello, E. K., Gordon, J. I., Secor, S. M., and Knight, R. (2010). Postprandial remodeling of the gut microbiota in Burmese pythons. *ISME J.* 4, 1375–1385. doi: 10.1038/ismej.2010.71

Crouch, L. I., Liberato, M. V., Urbanowicz, P. A., Baslé, A., Lamb, C. A., Stewart, C. J., et al. (2020). Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat. Commun.* 11:4017. doi: 10.1038/s41467-020-17847-5

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912. doi: 10.1093/bioinformatics/btq315

Depommier, C., Everard, A., Druart, C., Plovier, H., Van Hul, M., Vieira-Silva, S., et al. (2019). Supplementation with *Akkermansia muciniphila* in overweight and obese human volunteers: a proof-of-concept exploratory study. *Nat. Med.* 25, 1096–1103. doi: 10.1038/s41591-019-0495-2

Derrien, M., Belzer, C., and de Vos, W. M. (2017). *Akkermansia muciniphila* and its role in regulating host functions. *Microb. Pathog.* 106, 171–181. doi: 10.1016/j.micpath.2016.02.005

Derrien, M., Vaughan, E. E., Plugge, C. M., and de Vos, W. M. (2004). *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int. J. Syst. Evol. Microbiol.* 54, 1469–1476. doi: 10.1099/ijs.0.02873-0

Douglas, G. M., and Langille, M. G. I. (2019). Current and promising approaches to identify horizontal gene transfer events in metagenomes. *Genome Biol. Evol.* 11, 2750–2766. doi: 10.1093/gbe/evz184

Duarte, A. S., Correia, A., and Esteves, A. C. (2016). Bacterial collagenases - a review. *Crit. Rev. Microbiol.* 42, 106–126. doi: 10.3109/1040841X.2014.904270

Eaton, D. A. R. (2020). Toytree: a minimalist tree visualization and manipulation library for Python. *Methods Ecol. Evol.* 11, 187–191. doi: 10.1111/2041-210X.13313

Elhaik, E., Sabath, N., and Graur, D. (2006). The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol. Biol. Evol.* 23, 1–3. doi: 10.1093/molbev/msj006

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

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., et al. (2013). Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci. U. S. A.* 110, 9066–9071. doi: 10.1073/pnas.1219451110

Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., and Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* 3, 289–306. doi: 10.4161/gmic.19897

Geerlings, S. Y., Ouwerkerk, J. P., Koehorst, J. J., Ritari, J., Aalvink, S., Stecher, B., et al. (2021). Genomic convergence between *Akkermansia muciniphila* in different mammalian hosts. *BMC Microbiol.* 21:298. doi: 10.1186/s12866-021-02360-6

Glover, J. S., Ticer, T. D., and Engevik, M. A. (2022). Characterizing the mucin-degrading capacity of the human gut microbiota. *Sci. Rep.* 12:8456. doi: 10.1038/s41598-022-11819-z

Gomez, J. E., Kaufmann-Malaga, B. B., Wivagg, C. N., Kim, P. B., Silvis, M. R., Renedo, N., et al. (2017). Ribosomal mutations promote the evolution of antibiotic resistance in a multidrug environment. *Elife* 6:e20420. doi: 10.7554/eLife.20420

Grondin, J. A., Kwon, Y. H., Far, P. M., Haq, S., and Khan, W. I. (2020). Mucins in intestinal mucosal defense and inflammation: learning from clinical and experimental studies. *Front. Immunol.* 11:2054. doi: 10.3389/fimmu.2020.02054

Guo, X., Li, S., Zhang, J., Wu, F., Li, X., Wu, D., et al. (2017). Genome sequencing of 39 *Akkermansia muciniphila* isolates reveals its population structure, genomic and functional diversity, and global distribution in mammalian gut microbiotas. *BMC Genomics* 18:800. doi: 10.1186/s12864-017-4195-3

Hedlund, B. P. (2015). "Verrucomicrobia phyl. nov" in *Bergey's manual of systematics of archaea and bacteria*. eds. W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun and P. DeVoset al. (Chichester, UK: John Wiley & Sons, Ltd), 1–2.

Hedlund, B. P., and Derrien, M. (2015). "Akkermansiaceae fam. nov" in *Bergey's manual of systematics of archaea and bacteria*. eds. W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun and P. DeVoset al. (Chichester, UK: John Wiley & Sons, Ltd)

Hedlund, B. P., Yoon, J., and Kasai, H. (2015). "Rubritalea" in *Bergey's manual of systematics of archaea and bacteria*. eds. W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun and P. DeVoset al. (Chichester, UK: John Wiley & Sons, Ltd), 1–7.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* 464, 908–912. doi: 10.1038/nature08937

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Imam, S., Chen, Z., Roos, D. S., and Pohlschröder, M. (2011). Identification of surprisingly diverse type IV pili, across a broad range of gram-positive bacteria. *PLoS One* 6:e28919. doi: 10.1371/journal.pone.0028919

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

Jayaraman, V., Toledo-Patiño, S., Noda-García, L., and Laurino, P. (2022). Mechanisms of protein evolution. *Protein Sci.* 31:e4362. doi: 10.1002/pro.4362

Johansson, M. E. V., Larsson, J. M. H., and Hansson, G. C. (2011). The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4659–4665. doi: 10.1073/pnas.1006451107

Kamneva, O. K., Knight, S. J., Liberles, D. A., and Ward, N. L. (2012). Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol. Evol.* 4, 1375–1390. doi: 10.1093/gbe/evs113

Kanhere, A., and Vingron, M. (2009). Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol. Biol.* 9:9. doi: 10.1186/1471-2148-9-9

Karcher, N., Nigro, E., Punčochář, M., Blanco-Míguez, A., Ciciani, M., Manghi, P., et al. (2021). Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* 22:209. doi: 10.1186/s13059-021-02427-7

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformatics* 20, 1160–1166. doi: 10.1093/bib/bbx108

Kim, J.-S., Kang, S. W., Lee, J. H., Park, S.-H., and Lee, J.-S. (2022). The evolution and competitive strategies of *Akkermansia muciniphila* in gut. *Gut Microbes* 14:2025017. doi: 10.1080/19490976.2021.2025017

Kim, S., Shin, Y.-C., Kim, T.-Y., Kim, Y., Lee, Y.-S., Lee, S.-H., et al. (2021). Mucin degrader *Akkermansia muciniphila* accelerates intestinal stem cell-mediated epithelial development. *Gut Microbes* 13, 1–20. doi: 10.1080/19490976.2021.1892441

Kostopoulos, I., Aalvink, S., Kovatcheva-Datchary, P., Nijsse, B., Bäckhed, F., Knol, J., et al. (2021). A continuous Battle for host-derived glycans between a mucus specialist and a glycan generalist in vitro and in vivo. *Front. Microbiol.* 12:632454. doi: 10.3389/fmicb.2021.632454

Lang, T., Klasson, S., Larsson, E., Johansson, M. E. V., Hansson, G. C., and Samuelsson, T. (2016). Searching the evolutionary origin of epithelial mucus protein components-mucins and FCGBP. *Mol. Biol. Evol.* 33, 1921–1936. doi: 10.1093/molbev/msw066

Lawrence, J. G., and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843–1860. doi: 10.1093/genetics/143.4.1843

Levin, D., Raab, N., Pinto, Y., Rothschild, D., Zanir, G., Godneva, A., et al. (2021). Diversity and functional landscapes in the microbiota of animals in the wild. *Science* 372:eabb5352. doi: 10.1126/science.abb5352

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725

Li, W., Sun, J., Jing, Y., Zhao, J., Wu, Q., Liu, J., et al. (2022). Comparative genomics revealed wide intra-species genetic heterogeneity and lineage-specific genes of *Akkermansia muciniphila*. *Microbiol. Spectr.* 10:e0243921. doi: 10.1128/spectrum.02439-21

Ligthart, K., Belzer, C., de Vos, W. M., and Tytgat, H. L. P. (2020). Bridging bacteria and the gut: functional aspects of type IV pili. *Trends Microbiol.* 28, 340–348. doi: 10.1016/j.tim.2020.02.003

Lin, J., Zhou, D., Steitz, T. A., Polikanov, Y. S., and Gagnon, M. G. (2018). Ribosome-targeting antibiotics: modes of action, mechanisms of resistance, and implications for drug design. *Annu. Rev. Biochem.* 87, 451–478. doi: 10.1146/annurev-biochem-062917-011942

Lv, Q.-B., Li, S., Zhang, Y., Guo, R., Wang, Y.-C., Peng, Y., et al. (2022). A thousand metagenome-assembled genomes of Akkermansia reveal phylogroups and geographical and functional variations in the human gut. *Front. Cell. Infect. Microbiol.* 12:957439. doi: 10.3389/fcimb.2022.957439

Lv, Q.-B., Li, S.-H., Zhang, Y., Wang, Y.-C., Peng, Y.-Z., and Zhang, X.-X. (2020). A thousand metagenome-assembled genomes of *Akkermansia* reveal new phylogroups and geographical and functional variations in human gut. *BioRxiv*. doi: 10.1101/2020.09.10.292292

Ma, D., Cook, D. N., Alberti, M., Pon, N. G., Nikaido, H., and Hearst, J. E. (1995). Genes acrA and acrB encode a stress-induced efflux system of *Escherichia coli*. *Mol. Microbiol.* 16, 45–55. doi: 10.1111/j.1365-2958.1995.tb02390.x

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

N'Guessan, A., Brito, I. L., Serohijos, A. W. R., and Shapiro, B. J. (2021). Mobile gene sequence evolution within individual human gut microbiomes is better explained by gene-specific than host-specific selective pressures. *Genome Biol. Evol.* 13:evab142. doi: 10.1093/gbe/evab142

Ndongo, S., Armstrong, N., Raoult, D., and Fournier, P.-E. (2022). Reclassification of eight *Akkermansia muciniphila* strains and description of *Akkermansia massiliensis* sp. nov. and *Candidatus Akkermansia* timonensis, isolated from human feces. *Sci. Rep.* 12:21747. doi: 10.1038/s41598-022-25873-0

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

Nishimura, M., Yoshida, T., Shirouzu, M., Terada, T., Kuramitsu, S., Yokoyama, S., et al. (2004). Solution structure of ribosomal protein L16 from *Thermus thermophilus* HB8. *J. Mol. Biol.* 344, 1369–1383. doi: 10.1016/j.jmb.2004.10.011

Ojo, K. K., Tung, D., Luis, H., Bernardo, M., Leitao, J., and Roberts, M. C. (2004). Gram-positive *merA* gene in gram-negative oral and urine bacteria. *FEMS Microbiol. Lett.* 238, 411–416. doi: 10.1111/j.1574-6968.2004.tb09783.x

Ottman, N., Davids, M., Suarez-Diez, M., Boeren, S., Schaap, P. J., Martins Dos Santos, V. A. P., et al. (2017). Genome-scale model and omics analysis of metabolic capacities of *Akkermansia muciniphila* reveal a preferential mucin-degrading lifestyle. *Appl. Environ. Microbiol.* 83:e01014-17. doi: 10.1128/AEM.01014-17

Ottman, N., Huuskonen, L., Reunanen, J., Boeren, S., Klievink, J., Smidt, H., et al. (2016). Characterization of outer membrane proteome of *Akkermansia muciniphila* reveals sets of novel proteins exposed to the human intestine. *Front. Microbiol.* 7:1157. doi: 10.3389/fmicb.2016.01157

Ouwerkerk, J. P., Aalvink, S., Belzer, C., and de Vos, W. M. (2016a). *Akkermansia glycaniphila* sp. nov., an anaerobic mucin-degrading bacterium isolated from reticulated python faeces. *Int. J. Syst. Evol. Microbiol.* 66, 4614–4620. doi: 10.1099/ijsem.0.001399

Ouwerkerk, J. P., Tytgat, H. L. P., Elzinga, J., Koehorst, J., Van den Abbeele, P., Henrissat, B., et al. (2022). Comparative genomics and physiology of *Akkermansia muciniphila* isolates from human intestine reveal specialized mucosal adaptation. *Microorganisms* 10:1605. doi: 10.3390/microorganisms10081605

Ouwerkerk, J. P., van der Ark, K. C. H., Davids, M., Claassens, N. J., Finestra, T. R., de Vos, W. M., et al. (2016b). Adaptation of *Akkermansia muciniphila* to the oxic-anoxic interface of the mucus layer. *Appl. Environ. Microbiol.* 82, 6983–6993. doi: 10.1128/AEM.01641-16

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

Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R., and Stanton, C. (2022). Impact of antibiotics on the human microbiome and consequences for host health. *Microbiology* 11:e1260. doi: 10.1002/mbo3.1260

Patra, A. K., and Yu, Z. (2022). Genomic insights into the distribution of peptidases and proteolytic capacity among Prevotella and Paraprevotella species. *Microbiol. Spectr.* 10:e0218521. doi: 10.1128/spectrum.02185-21

Png, C. W., Lindén, S. K., Gilshenan, K. S., Zoetendal, E. G., McSweeney, C. S., Sly, L. I., et al. (2010). Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. *Am. J. Gastroenterol.* 105, 2420–2428. doi: 10.1038/ajg.2010.281

Presley, L. L., Wei, B., Braun, J., and Borneman, J. (2010). Bacteria associated with immunoregulatory cells in mice. *Appl. Environ. Microbiol.* 76, 936–941. doi: 10.1128/AEM.01561-09

Qi, M., Nelson, K. E., Daugherty, S. C., Nelson, W. C., Hance, I. R., Morrison, M., et al. (2005). Novel molecular features of the fibrolytic intestinal bacterium *Fibrobacter intestinalis* not shared with *Fibrobacter succinogenes* as determined by suppressive subtractive hybridization. *J. Bacteriol.* 187, 3739–3751. doi: 10.1128/JB.187.11.3739-3751.2005

Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Comput. Biol.* 11:e1004095. doi: 10.1371/journal.pcbi.1004095

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., and Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* 46, D624–D632. doi: 10.1093/nar/gkx1134

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

Sauvaitre, T., Etienne-Mesmin, L., Sivignon, A., Mosoni, P., Courtin, C. M., Van de Wiele, T., et al. (2021). Tripartite relationship between gut microbiota, intestinal mucus and dietary fibers: towards preventive strategies against enteric infections. *FEMS Microbiol. Rev.* 45:fuaa052. doi: 10.1093/femsre/fuaa052

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

Shin, N.-R., Lee, J.-C., Lee, H.-Y., Kim, M.-S., Whon, T. W., Lee, M.-S., et al. (2014). An increase in the *Akkermansia* spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* 63, 727–735. doi: 10.1136/gutjnl-2012-303839

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C.-H., Westover, B. P., Weatherford, J., et al. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* 307, 1955–1959. doi: 10.1126/science.1109051

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16, 472–482. doi: 10.1038/nrg3962

Spero, M. A., Aylward, F. O., Currie, C. R., and Donohue, T. J. (2015). Phylogenomic analysis and predicted physiological role of the proton-translocating NADH:quinone oxidoreductase (complex I) across bacteria. *mBio* 6:e00389-15. doi: 10.1128/mBio.00389-15

Stajich, J. E. (2007). An introduction to bioperl. *Methods Mol. Biol.* 406, 535–548. doi: 10.1007/978-1-59745-535-0_26

Stam, M., Lelièvre, P., Hoebeke, M., Corre, E., Barbeyron, T., and Michel, G. (2023). SulfAtlas, the sulfatase database: state of the art and new developments. *Nucleic Acids Res.* 51, D647–D653. doi: 10.1093/nar/gkac977

Sun, J., Deng, Z., and Yan, A. (2014). Bacterial multidrug efflux pumps: mechanisms, physiology and pharmacological exploitations. *Biochem. Biophys. Res. Commun.* 453, 254–267. doi: 10.1016/j.bbrc.2014.05.090

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Takeuchi, N., Wolf, Y. I., Makarova, K. S., and Koonin, E. V. (2012). Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.* 194, 1216–1225. doi: 10.1128/JB.06521-11

Tarracchini, C., Milani, C., Lugli, G. A., Mancabelli, L., Fontana, F., Alessandri, G., et al. (2021). Phylogenomic disentangling of the *Bifidobacterium longum* subsp. infantis taxon. *Microb. Genom.* 7:000609. doi: 10.1099/mgen.0.000609

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

Treangen, T. J., and Rocha, E. P. C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7:e1001284. doi: 10.1371/journal.pgen.1001284

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.* 804, 281–295. doi: 10.1007/978-1-61779-361-5_15

Van Oss, S. B., and Carvunis, A.-R. (2019). De novo gene birth. *PLoS Genet.* 15:e1008160. doi: 10.1371/journal.pgen.1008160

van Passel, M. W. J., Kant, R., Zoetendal, E. G., Plugge, C. M., Derrien, M., Malfatti, S. A., et al. (2011). The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One* 6:e16876. doi: 10.1371/journal.pone.0016876

Vidal-Veuthey, B., González, D., and Cárdenas, J. P. (2022). Role of microbial secreted proteins in gut microbiota-host interactions. *Front. Cell. Infect. Microbiol.* 12:964710. doi: 10.3389/fcimb.2022.964710

Vinderola, G., Sanders, M. E., Salminen, S., and Szajewska, H. (2022). Postbiotics: the concept and their use in healthy populations. *Front. Nutr.* 9:1002213. doi: 10.3389/fnut.2022.1002213

Wallace, R. J., and McKain, N. (1997). Peptidase activity of human colonic bacteria. *Anaerobe* 3, 251–257. doi: 10.1006/anae.1997.0080

Wolska, K. I. (2003). Horizontal DNA transfer between bacteria in the environment. *Acta Microbiol. Pol.* 52, 233–243.

Xing, J., Li, X., Sun, Y., Zhao, J., Miao, S., Xiong, Q., et al. (2019). Comparative genomic and functional analysis of *Akkermansia muciniphila* and closely related species. *Genes Genomics* 41, 1253–1264. doi: 10.1007/s13258-019-00855-1

Xu, W., Yang, W., Wang, Y., Wang, M., and Zhang, M. (2020). Structural and biochemical analyses of β-N-acetylhexosaminidase Am0868 from *Akkermansia muciniphila* involved in mucin degradation. *Biochem. Biophys. Res. Commun.* 529, 876–881. doi: 10.1016/j.bbrc.2020.06.116

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

Youngblut, N. D., de la Cuesta-Zuluaga, J., Reischer, G. H., Dauser, S., Schuster, N., Walzer, C., et al. (2020). Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. *mSystems* 5:e01045-20. doi: 10.1128/mSystems.01045-20

Zaneveld, J. R. (2011). *Habitat adaptation and genome evolution in the gut microbiome*. Available at: https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/d217qp592

Zhang, M., Chen, H., Liu, L., Xu, L., Wang, X., Chang, L., et al. (2020). The changes in the frog gut microbiome and its putative oxygen-related phenotypes accompanying the development of gastrointestinal complexity and dietary shift. *Front. Microbiol.* 11:162. doi: 10.3389/fmicb.2020.00162

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., et al. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 46, W95–W101. doi: 10.1093/nar/gky418

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

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 10 October 2023
DOI 10.3389/fmicb.2023.1278268

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Ram Nageena Singh,
South Dakota School of Mines and Technology, United States
Sonia Dávila-Ramos,
Autonomous University of the State of Morelos, Mexico

*CORRESPONDENCE
Maojun Zhang
✉ zhangmaojun@icdc.cn

RECEIVED 16 August 2023
ACCEPTED 27 September 2023
PUBLISHED 10 October 2023

CITATION
Zhou G, Gu Y, Wang H, Chen X, Zhang X, Shao Z, Yan X, Zhang J and Zhang M (2023) Genomic diversity and taxonomic marker for *Arcobacter* species.
Front. Microbiol. 14:1278268.
doi: 10.3389/fmicb.2023.1278268

COPYRIGHT
© 2023 Zhou, Gu, Wang, Chen, Zhang, Shao, Yan, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Genomic diversity and taxonomic marker for *Arcobacter* species

Guilan Zhou, Yixin Gu, Hairui Wang, Xiaoli Chen, Xin Zhang, Zhujun Shao, Xiaomei Yan, Jianzhong Zhang and Maojun Zhang*

State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China

Arcobacter was recognized as an emerging enteropathogen and controversies regarding its classification persisted. This study aimed to reevaluate the taxonomy of *Arcobacter* utilizing the 16S rRNA gene, 23S rRNA gene, single-copy orthologous genes, as well as genomic indices such as Average Nucleotide Identity (ANI) and *in silico* DNA–DNA hybridization (*isDDH*). The taxonomy of this genus was reevaluated in this study using multiple indices with a dataset of 371 genomes comprising 34 known species and 14 potentially new species. Good discrimination could be achieved only in some species but not for the species with higher sequence similarity using the comparisons of the 16S rRNA gene and 23S rRNA gene sequences. A high-accuracy phylogenomic approach for *Arcobacter* was established using 84 single-copy orthologous genes obtained through various bioinformatics methods. One marker gene (gene711), which was found to possess the same distinguishing ability as ANI, *isDDH,* and single-copy orthologous methods, was identified as a reliable locus for inferring the phylogeny of the genus. The effective species classification was achieved by employing gene711 with a sequence similarity exceeding 96%, even for species like *A. cloacae*, *A. lanthieri*, and *A. skirrowii*, which exhibited ambiguous classification using ANI and *isDDH*. Additionally, excellent subspecies categorizing among *A. cryaerophilus* could be distinguished using gene711. In conclusion, this framework strategy had the potential advantage of developing rapid species identification, particularly for highly variable species, providing a novel insight into the behavior and characteristics of *Arcobacter*.

KEYWORDS

***Arcobacter*, genome sequencing, taxonomy, ANI, *isDDH*, reliable marker gene**

Introduction

Arcobacter has gained increasing significance in recent years, as its members are now recognized as emerging enteropathogens and potential zoonotic agents (Ho et al., 2006). The *Arcobacter* genus belongs to the *Campylobacteraceae* family, which includes other genera: *Campylobacter*, *Helicobacter*, *Sulfurospirillum*, and others (On, 2001). Initially classified within the *Campylobacter* genus, it was in 1991 that the *Arcobacter* genus was recognized as distinct and designated as a separate genus within the *Campylobacteraceae* family (Vandamme et al., 1991). *Arcobacter* was generally described as possessing differentiated abilities from *Campylobacter*, namely the ability to grow in aerobic conditions and at temperatures between 15 and 30°C (Vandamme et al., 1992); however, this principle has been changed by the increased number of new species. Nowadays, *Arcobacter* species inhabit a wide range of ecological niches, encompassing diverse environments such as marine environments, wastewater and drinking

water systems, animal feces, plants, and even oil fields, among others (Van Driessche et al., 2005; Collado and Figueras, 2011; Rathlavath et al., 2017; On et al., 2021; Pascual et al., 2023). Some *Arcobacter* species have been detected in or isolated from the stools of patients with and without diarrhea, occasionally being associated with conditions such as bacteremia, endocarditis, and peritonitis (Vandenberg et al., 2004; Ho et al., 2006; Van den Abeele et al., 2014; Isidro et al., 2020). Furthermore, it is crucial to acknowledge that the actual prevalence of *Arcobacter* species may be underestimated due to the constraints imposed by current detection and identification methods (Hanel et al., 2016). Currently, the *Arcobacter* genus consists of 34 species with validly published and accurately designated names[1] (Pascual et al., 2023). In previous studies, the similarity of the 16S rRNA gene was considered a decisive characteristic for a taxonomic assignation at the genus level (Roth et al., 2003; Clarridge, 2004). However, misclassifications have been observed when comparing closely related species based solely on phylogenetic analysis of the 16S rRNA gene, attributed to their high sequence similarities. Debruyne et al. (2010) demonstrated that the *hsp60* gene provided higher resolution than the 16S rRNA gene in closely related species. Nonetheless, caution should be exercised when utilizing this gene alone for species-level identification within taxa characterized by high genomic diversity. Subsequently, a multilocus sequence analysis (MLSA) that relies on multiple conserved molecular markers (*atpA*, *atpD*, *dnaA*, *dnaJ*, *dnaK*, *ftsZ*, *gyrA*, *hsp60*, *radA*, *recA*, *rpoB*, *rpoD*, and *tsf*) have been investigated to differentiate species better and determine their phylogenetic relationships (Debruyne et al., 2010; Perez-Cataluna et al., 2018b). However, irrespective of the methodology employed, the identification of uncommon *Arcobacter* species remains challenging. In a taxonomy study conducted by Perez-Cataluna et al. (2018b), several approaches, including Average Nucleotide Identity (ANI), *in silico* DNA–DNA Hybridization (*isDDH*), Average Amino-acid Identity, Percentage of Conserved Proteins, and Relative Synonymous Codon Usage were employed to address this issue. The study suggested that the current *Arcobacter* genus should be divided into at least seven different genera: *Arcobacter*, *Aliarcobacter*, *Haloarcobacter*, *Pseudoarcobacter*, *Poseidonibacter*, *Malacobacter*, and Candidate '*Arcomarinus*' gen. Nov (Perez-Cataluna et al., 2018b). However, On et al. (2020) revealed that the *Arcobacter* genus displayed relatively homogenous, and phylogenetic analyzes clearly distinguished this group from other *Epsilonproteobacteria* and showed that any of the measures used did not support the genomic distinction of the genera proposed by Perez-Cataluna et al. It is noteworthy that the proposal put forward by Perez-Cataluna et al. has not received approval from the International Committee on Systematics of Prokaryotes taxonomy subcommittee on *Campylobacter* or nor has it been validated in the International Journal of Systematic and Evolutionary Microbiology (On et al., 2021).

The field of prokaryotic systematics has been dramatically changed by the emergence of genome sequencing, resulting in significant advancements in various aspects, including species identification, functional characterization for taxonomic delineation, and the elucidation of phylogenetic relationships at higher taxonomic levels (Whitman, 2015). Moreover, with the advancement of detection methods, the number of *Arcobacter* strains is increasing, leading to the gradual identification of new *Arcobacter* species. Consequently, this progress poses a challenge in effectively classifying these species, thereby introducing increased difficulties in taxonomy. Incorporating genomics into taxonomy appears to be a promising development, enhancing credibility by offering reproducible, reliable, and highly informative methods to infer phylogenetic relationships among prokaryotes while avoiding unreliable approaches and subjective, difficult-to-replicate data. Within this modern taxonomy context, the objective of this study was to reassess the taxonomy of both known and newly identified *Arcobacter* species by using 16S rRNA gene, 23S rRNA gene, the whole genome sequences, and the derived genomic analysis, providing valuable insights into the taxonomic investigation of *Arcobacter*. We also evaluated the efficacy of various genome-based phylogenetic tools in discriminating between different *Arcobacter* species.

Materials and methods

Bacterial strains

In this study, 371 *Arcobacter* genomes were used, out of which 172 were obtained from strains sequenced by our laboratory or collaborating institutions. The isolation, cultivation, genomic DNA extraction, and sequencing of these strains were described in previous publications (Wang et al., 2021; Ma et al., 2022; Zhou et al., 2022). Furthermore, genomes of *Arcobacter* identified at the species level were investigated, 172 of which were obtained in our earlier studies (70 *A. butzleri*, 81 *A. cryaerophilus*, 19 *A. skirrowii*, and 2 *A. lacus*), and the others from the public databases. checkM software (Parks et al., 2015) was used to assess genomic contamination and completeness, resulting in contamination <4.67% (CNAS04, CNAC065) and completeness >96.34% (CNAB027). The 371 genomes were annotated with a local installation of Prokka v1.14.6 (Seemann, 2014) with the prediction tools Prodigal v2.6.3 (Hyatt et al., 2010) and ARAGORN v1.2.41 (Laslett and Canback, 2004). The prediction tool barrnap v0.9[2] included in Prokka v1.14.6 was used to annotate rRNA genes. The characteristics of each genome (i.e., N50, number of contigs, G + C content) were obtained using in-house scripts.

Downloading of publicly available genomes

All 34 valid species included in the *Arcobacter* genus have been studied. They were represented by 199 genomes and 17 potentially new species genomes (Supplementary Table S1). All genome sequences identified as *Arcobacter* were downloaded from the National Center for Biotechnology Information (NCBI) and Bacterial and Viral Bioinformatics Resource Center (BV-BRC) public database on January 2023. All publicly available assemblies were subjected to quality control by Quast software (Gurevich et al., 2013). Firstly, genomic sequences identified as "poor" were excluded from the

1 https://lpsn.dsmz.de/genus/arcobacter

2 https://github.com/tseemann/barrnap

analysis based on the sequencing quality. Secondly, genomes that did not meet the criteria for genome size and GC content were filtered out according to the genomic characteristics of *Arcobacter*. Additionally, only genomes with a scaffold count of less than 200 were included to ensure the reliability of the analysis results. Finally, the obtained genomes underwent species identification using the GTDB v2.3.2 software (Chaumeil et al., 2019), and only the genomes identified as *Arcobacter* were included in the analysis. A total of 199 *Arcobacter* genomes were included in the study, comprising 34 named *Arcobacter* species and 14 unclassified *Arcobacter* species, as shown in Supplementary Table S1.

Analysis of ribosomal genes

The 16S rRNA gene and 23S rRNA gene sequences were extracted from the genome assemblies using barrnap v0.9, producing a gff file of rRNA gene locations in the genome assemblies. The gff files were combined with the bedtools (Quinlan and Hall, 2010), fastaFromBed, to extract the 16S rRNA and 23S rRNA gene sequences from the genome assemblies. Genes sequences were aligned using MAFFT v7.490 software (Katoh and Standley, 2013). The genomes containing the complete 16S rRNA gene and 23S rRNA gene were selected, and the corresponding sequences were extracted and aligned to construct a Neighbor-Joining (NJ) phylogenetic tree with a bootstrap value of 1,000. Additionally, pairwise sequence comparisons were performed using MAFFT v7.490 software (Katoh and Standley, 2013) to determine sequence alignments and assess the similarity between pairs of sequences.

Analysis of ANI and *isDDH*

Pairwise ANI values were calculated for all genomes using pyani v0.2 software (module ANIb), accessible at https://github.com/widdowquinn/pyani. The Genome-to-Genome Distance Calculator (GGDC) web service was used to report *isDDH* for the accurate delineation of prokaryotic subspecies and to calculate differences in G + C genomic content.[3] Analysis was performed using "Formula 2," as recommended by the GGDC authors, which allows for *isDDH* estimation independent of genome lengths, making it suitable for incomplete genomes. A matrix with ANI values across all genomes was visualized using the pheatmap package, and an in-house script was used to generate a clustering dendrogram based on the ANI matrix.

Identification of single-copy orthologous genes and marker gene

The OrthoFinder v2.5.4 software (Emms and Kelly, 2019) was employed to perform a homology analysis on the 371 *Arcobacter* genomes, identifying single-copy orthologous genes. The software parameters used were -S blast, −M msa, −T raxml. The EasyTree.py script[4] was used to extract all single-copy orthologous genes from each genome. The genes were aligned using the MAFFT v7.490 software, and an ML tree (data not shown) was constructed by concatenating and coalescing these genes using the raxmlHPC v8.2.12 software (Stamatakis, 2014) and MEGA 7 (Kumar et al., 2016) software, with a bootstrap value of 1,000. The resulting tree was annotated using the table2itol package and visualized in iTOL.[5]

Results

Genomic characteristics of the *Arcobacter*

A total of 371 high-quality sequenced and assembled genomes of *Arcobacter* were obtained through genome quality control, and a comprehensive analysis was conducted on 371 genomes. All 34 species currently included in the *Arcobacter* and 14 candidate species have been investigated in the present study. The scaffolds obtained and the N50 values complied with the proposed minimal standards for using genomes in taxonomic studies (Chun et al., 2018). Genome assemblies had 1 to 166 contigs. The genome sizes and GC contents displayed significant variations across different *Arcobacter* species. The genome size ranged from 1.68 Mb for *A. skirrowii* CNAS13 to 3.57 Mb for *A. lekithochrous* CP054052. The genome size of *A. skirrowii* was generally smaller than that of other *Arcobacter* species. In comparison, the genome size of *A. lekitochrous* was generally larger than that of other *Arcobacter* species. The G + C content ranged from 26.08% in *A. molluscorum* NXFY00000000 to 31.00% in *Arcobacter* spp. JAIFNA000000000, as shown in Supplementary Table S1.

Phylogenetic of ribosomal genes

The size of the 16S rRNA gene in 34 type strains of *Arcobacter* species ranged from 1,512 to 1,516 bp, with sequence similarities ranging from 91.97% (between *A. cryaerophilus* and *A. bivalviorum*) to 99.93% (between *A. butzleri* and *A. lacus*). Similarly, the size of the 23S rRNA gene varied from 2,873 to 3,026 bp, with sequence similarities ranging from 86.72% (between *A. vandammei* and *A. pacificus*) to 99.72% (between *A. butzleri* and *A. lacus*). Detailed results can be found in Tables 1, 2 and Supplementary Table S2. The phylogenetic trees constructed based on the 16S rRNA gene and 23S rRNA gene of the type strains were presented in Figure 1. It was noteworthy that there were certain variations observed in the phylogenetic trees constructed using different sequence datasets. Of the 371 *Arcobacter* genomes analyzed, 281 were selected for analysis due to the near-full length of the 16S rRNA gene and 23S rRNA gene. The size of the 16S rRNA gene ranged from 1,306 to 1,517 bp, almost all of which were around 1,514 bp, except VBUD00000000, VBUC00000000, NXGJ00000000, SZACF0142G, SZACF1311G, and SZACF1324G. Similarly, the size of the 23S rRNA gene ranged from 2,607 to 3,030 bp, most of which were around 2,907 bp. The similarities in the 16S rRNA gene sequences among different *Arcobacter* species

3 Available at http://ggdc.dsmz.de/ggdc.php.

4 https://github.com/dongwei1220/EasySpeciesTree

5 https://itol.embl.de/

10.3389/fmicb.2023.1278268

TABLE 1 16S rRNA gene and 23S rRNA gene sizes and the start and end positions of gene711 in the genome of 34 *Arcobacter* species type strains.

Species	GenBank Accession (Type Strain)	16S rRNA	23S rRNA	gene711	gene711_ Start	gene711_ End	Note*
Arcobacter acticola	CP042652	1,514	2,949	597	2,885,771	2,886,367	
Arcobacter anaerophilus	CP041070	1,514	2,878	618	2,913,613	2,914,230	
Arcobacter antarcticus	RCWF00000000	1,514	2,909	585	112,316	112,900	fasta56
Arcobacter aquimarinus	CP030944	1,515	2,908	585	2,431,557	2,432,141	
Arcobacter arenosus	VANU00000000	1,515	2,873	609	378,208	378,816	fasta12
Arcobacter bivalviorum	CP031217	1,515	2,905	654	2,582,303	2,582,956	
Arcobacter butzleri	CP000361	1,514	2,907	591	2,262,773	2,263,363	
Arcobacter caeni	MUXE00000000	1,514	2,906	618	15,779	16,396	fasta57
Arcobacter cibarius	CP054051	1,515	2,908	585	1,968,875	1,969,459	
Arcobacter cloacae	CP053834	1,515	2,908	588	2,533,454	2,534,041	
Arcobacter cryaerophilus	CP032824	1,515	2,907	600	1,897,718	1,898,317	
Arcobacter defluvii	CP053835	1,515	2,906	588	2,904,720	2,905,307	
Arcobacter ebronensis	CP053836	1,515	2,879	615	3,035,147	3,035,761	
Arcobacter ellisii	CP032097	1,515	2,906	600	2,702,365	2,702,964	
Arcobacter faecis	CP053838	1,515	2,907	588	1,991,920	1,992,507	
Arcobacter halophilus	CP031218	1,514	2,908	576	2,741,754	2,742,329	
Arcobacter lacus	MUXF00000000	1,514	2,907	603	98,969	99,571	fasta16
Arcobacter lanthieri	CP053839	1,515	2,907	603	122,801	123,403	
Arcobacter lekithochrous	CP054052	1,514	2,908	588	3,484,870	3,485,457	
Arcobacter marinus	CP032101	1,512	2,915	576	2,842,957	2,843,532	
Arcobacter molluscorum	CP032098	1,514	2,908	576	2,724,302	2,724,877	
Arcobacter mytili	CP031220	1,514	2,908	573	2,795,077	2,795,649	
Arcobacter nitrofigilis	CP001999	1,514	2,911	558	3,065,193	3,065,750	
Arcobacter pacificus	CP035928	1,514	2,921	597	2,556,007	2,556,603	
Arcobacter parvus	CP019070	1,514	2,909	585	2,800,450	2,801,034	
Arcobacter porcinus	CP036246	1,515	2,907	615	90,336	90,950	
Arcobacter skirrowii	CP032099	1,514	2,909	618	126,189	126,806	
Arcobacter suis	CP032100	1,515	2,906	615	2,530,252	2,530,866	
Arcobacter thereius	CP035926	1,515	2,907	615	91,036	91,650	
Arcobacter trophiarum	CP031367	1,515	2,908	597	103,102	103,698	
Arcobacter vandammei	JADKPZ000000000	1,514	3,026	588	58,759	59,346	fasta23
Arcobacter venerupis	CP053840	1,514	2,906	609	3,051,151	3,051,759	
Arcobacter vitoriensis	PDKB00000000	1,515	2,907	603	71,009	71,611	fasta5
Arcobacter roscoffensis	CP100595	1,514	2,908	588	3,076,989	3,077,576	

Arcobacter antarcticus, *Arcobacter arenosus*, *Arcobacter caeni*, *Arcobacter lacus*, *Arcobacter vandammei*, and *Arcobacter vitoriensis* are draft genomes, and the start and end positions of gene711 are scaffold56, scaffold12, scaffold57, scaffold16, scaffold23, and scaffold5, respectively.

(all the 34 species currently included in the genus and the 14 new candidate species) showed a wide range of values (Table 2; Supplementary Table S2). Similarities ranged from 89.10% (between *A. anaerophilus*_CP041070 and *A.* spp_CP041403) to 100% (between *A. butzleri* and *A. lacus*). Notably, the similarity of the 16S rRNA gene between some *Arcobacter* species reached or even exceeded the similarity within species, such as *A. cloacae* and *A. ellisii*, *A. cryaerophilus* and *A. skirrowii*, *A. lacus* and *A. butzleri* and others. The differences in 23S rRNA gene sequences among different *Arcobacter* species were greater compared to the 16S rRNA gene sequences, with sequence similarities ranging from 83.60% (between *A. vandammei* and *A. marinus*) to 99.76% (between *A. butzleri* and *A. lacus*). However, the similarity of 23S rRNA gene sequences among some species still exceeded the similarity within species such as *A. cryaerophilus* and *A. skirrowii*, *A. lacus* and *A. butzleri*. Figure 2 and Supplementary Figure S1 illustrate the phylogenetic relationships of the 16S rRNA gene and 23S rRNA gene among the presently described species. Although these two phylogenetic trees showed high topological similarity, neither of them effectively distinguished species within the *Arcobacter* genus, as evidenced by the inability to

TABLE 2 Intraspecies and interspecies similarity of 16S rRNA gene, 23S rRNA gene, ANI, and gene711 of *Arcobacter*.

	16S rRNA gene in intraspecies	16S rRNA gene in interspecies (Closest related)	23S rRNA gene in intraspecies	23S rRNA gene in interspecies (Closest related)	ANI in intraspecies	ANI in interspecies (Closest related)	gene711 in intraspecies	gene711 in interspecies (Closest related)
A. acticola	100%	*A. venerupis* (98.40%)	100%	*A. caeni* (96.91%)	100.00%	*A. caeni* (83.06%)	100.00%	*A. cloacae* (77.22%)
A. anaerophilus	100%	*A. ebronensis* (96.60%)	100%	*A. ebronensis* (96.25%)	100.00%	*A. ebronensis* (79.13%)	100.00%	*A. ebronensis* (69.26%)
A. antarcticus	100%	*A. parvus* (98.50%)	100%	*A. parvus* (97.42%)	100.00%	*A. parvus* (84.04%)	100.00%	*A.* spp_CAJWWD000000000 (77.44%)
A. aquimarinus	100%	*A. defluvii* (99.41%)	100%	*A. cloacae* (99.42%)	97.05%	*A. cloacae* (93.82%)	100.00%	*A. cloacae* (93.88%)
A. arenosus	100%	*A.* spp_PDKF00000000 (96.70%)	100%	*A. ebronensis* (95.59%)	100.00%	*A. bivalviorum* (78.01%)	100.00%	*A. ebronensis* (60.89%)
A. bivalviorum	99.60%	*A.* spp_PDJU00000000 (99.27%)	99.93%	*A.* spp_PDKF00000000 (97.04%)	96.52%	*A.* spp_PDJU00000000 (85.66%)	97.71%	*A.* spp_PDJU00000000 (84.40%)
A. butzleri	99.93%	*A. lacus* (100%)	99.73%	*A. lacus* (99.76%)	96.75%	*A. lacus* (94.60%)	96.62%	*A. lacus* (84.41%)
A. caeni	100%	*A. venerupis* (99.47%)	100%	*A. suis* (99.04%)	100.00%	*A. suis* (87.86%)	100.00%	*A. suis* (86.41%)
A. cibarius	99.92%	*A. cryaerophilus* (99.01%)	100%	*A. faecis* (98.35%)	98.85%	*A. faecis* (84.79%)	98.97%	*A. vandammei* (84.69%)
A. cloacae	99.80%	*A. ellisii* (99.80%)	99.90%	*A. aquimarinus* (99.42%)	95.61%	*A. aquimarinus* (93.82%)	96.43%	*A. aquimarinus* (93.88%)
A. cryaerophilus	98.95%	*A. skirrowii* (99.51%)	93.05%	*A. skirrowii* (99.11%)	92.32%	*A. trophiarum* (85.78%)	92.33%	*A. trophiarum* (88.83%)
A. cryaerophilus-I	99.41%	*A. cryaerophilus-II* (100%)	93.05%	*A. cryaerophilus-2* (99.76%)	96.00%	*A. cryaerophilus-II* (96.26%)	95.33%	*A. cryaerophilus-II* (96.50%)
A. cryaerophilus-II	99.34%	*A. cryaerophilus-I* (100%)	99.48%	*A. cryaerophilus-3* (99.90%)	96.33%	*A. cryaerophilus-I* (96.26%)	98.17%	*A. cryaerophilus-I* (96.50%)
A. cryaerophilus-III	100%	*A. cryaerophilus-II* (99.93%)	100%	*A. cryaerophilus-2* (99.90%)	100.00%	*A. cryaerophilus-IV* (94.26%)	100.00%	*A. cryaerophilus-IV* (95.83%)
A. cryaerophilus-IV	100%	*A. cryaerophilus-III* (99.54%)	100%	*A. cryaerophilus-3* (99.31%)	97.78%	*A. cryaerophilus-III* (94.34%)	99.00%	*A. cryaerophilus-III* (95.83%)
A. defluvii	100%	*A. cloacae/A.aquimarinus* (99.41%)	100%	*A. ellisii* (99.38%)	100.00%	*A. aquimarinus* (84.46%)	100.00%	*A. cloacae* (85.67%)
A. ebronensis	99.47%	*A. bivalviorum* (96.70%)	99.50%	*A. anaerophilus* (96.25%)	96.10%	*A. anaerophilus* (79.13%)	96.75%	*A. anaerophilus* (69.26%)
A. ellisii	100%	*A. cloacae* (99.80%)	100%	*A. defluvii* (99.38%)	96.42%	*A. defluvii* (87.58%)	98.00%	*A. defluvii* (85.67%)
A. faecis	100%	*A. lanthieri* (99.14%)	100%	*A. cibarius* (98.35%)	99.80%	*A. vandammei* (85.20%)	100.00%	*A. vandammei* (85.03%)
A. halophilus	100%	*A. marinus* (96.63%)	100%	*A. molluscorum* (96.87%)	99.97%	*A. marinus* (86.55%)	100.00%	*A. marinus* (86.11%)
A. lacus	99.93%	*A. butzleri* (100%)	99.97%	*A. butzleri* (99.76%)	100.00%	*A. butzleri* (94.56%)	99.50%	*A. butzleri* (84.41%)
A. lanthieri	99.80%	*A. vitoriensis* (99.27%)	99.62%	*A. vitoriensis* (98.87%)	95.67%	*A. vitoriensis* (87.69%)	96.19%	*A. vitoriensis* (87.40%)
A. lekithochrous	100%	*A. roscoffensis* (96.51%)	100%	*A. roscoffensis* (96.60%)	98.33%	*A. roscoffensis* (79.24%)	98.64%	*A. roscoffensis* (74.66%)
A. marinus	99.60%	*A. molluscorum* (97.69%)	93.81%	*A. halophilus* (96.33%)	95.26%	*A. halophilus* (86.48%)	94.79%	*A. halophilus* (86.11%)

(Continued)

10.3389/fmicb.2023.1278268

TABLE 2 (Continued)

	16S rRNA gene in intraspecies	16S rRNA gene in interspecies (Closest related)	23S rRNA gene in intraspecies	23S rRNA gene in interspecies (Closest related)	ANI in intraspecies	ANI in interspecies (Closest related)	gene711 in intraspecies	gene711 in interspecies (Closest related)
A. molluscorum	100%	*A.* spp_PDKC00000000 (98.55%)	100%	*A.* spp_PDKG00000000 (98.56%)	100.00%	*A.* spp_PDKG00000000 (90.23%)	100.00%	*A.* spp_PDKG00000000 (88.54%)
A. mytili	99.93%	*A.* spp_PDKG00000000 (95.24%)	99.69%	*A. halophilus* (94.98%)	98.97%	*A. halophilus* (80.80%)	100.00%	*A. molluscorum* (74.44%)
A. nitrofigilis	100%	*A.* spp_PDJV00000000 (99.21%)	100%	*A.* spp_PDJV00000000 (99.28%)	100.00%	*A.* spp_PDJV00000000 (91.83%)	100.00%	*A.* spp_PDJV00000000 (86.38%)
A. pacificus	100%	*A. roscoffensis* (96.30%)	100%	*A. arenosus* (93.37%)	99.86%	*A. cloacae* (78.88%)	100.00%	*A. parvus* (72.26%)
A. parvus	99.87%	*A. antarcticus* (98.55%)	99.38%	*A. antarcticus* (97.42%)	98.09%	*A. antarcticus* (84.27%)	99.49%	*A.* spp_CAJWWD000000000/*A. spp*_NYWO00000000 (78.46%)
A. porcinus	100%	*A. thereius* (99.01%)	99.90%	*A. thereius* (99.42%)	98.22%	*A. thereius* (93.51%)	99.84%	*A. thereius* (92.52%)
A. skirrowii	99.14%	*A. cryaerophilus* (99.51%)	92.85%	*A. cryaerophilus* (99.11%)	94.83%	*A. cryaerophilus* (94.83%)	96.12%	*A. thereius* (78.69%)
A. suis	99.93%	*A. cloacae* (99.08%)	100%	*A. venerupis* (99.35%)	99.92%	*A. caeni* (87.94%)	100.00%	*A. caeni* (86.41%)
A. thereius	100%	*A. porcinus* (99.01%)	100%	*A. porcinus* (99.42%)	98.63%	*A. porcinus* (93.49%)	99.35%	*A. porcinus* (92.52%)
A. trophiarum	99.93%	*A. cryaerophilus* (98.94%)	100%	*A. porcinus* (98.62%)	99.74%	*A. cryaerophilus* (85.82%)	100.00%	*A. cryaerophilus* (88.83%)
A. vandammei	100%	*A. lanthieri* (98.42%)	100%	*A. skirrowii* (96.80%)	100.00%	*A. faecis* (85.25%)	100.00%	*A. faecis* (85.03%)
A. venerupis	100%	*A. caeni* (99.47%)	100%	*A. suis* (99.35%)	100.00%	*A. suis* (86.65%)	100.00%	*A. caeni* (79.94%)
A. vitoriensis	99.93%	*A. lanthieri* (99.27%)	100%	*A. lanthieri* (98.87%)	97.31%	*A. lanthieri* (87.19%)	98.84%	*A. lanthieri* (87.40%)
Arcobacter roscoffensis	100%	*A. pacificus* (96.30%)	100%	*A. lekithochrous* (96.60%)	100.00%	*A. lekithochrous* (79.24%)	100.00%	*A. lekithochrous* (74.66%)

The highlighted content in the table indicates species with ambiguous classifications when employing various indices for species identification.

10.3389/fmicb.2023.1278268

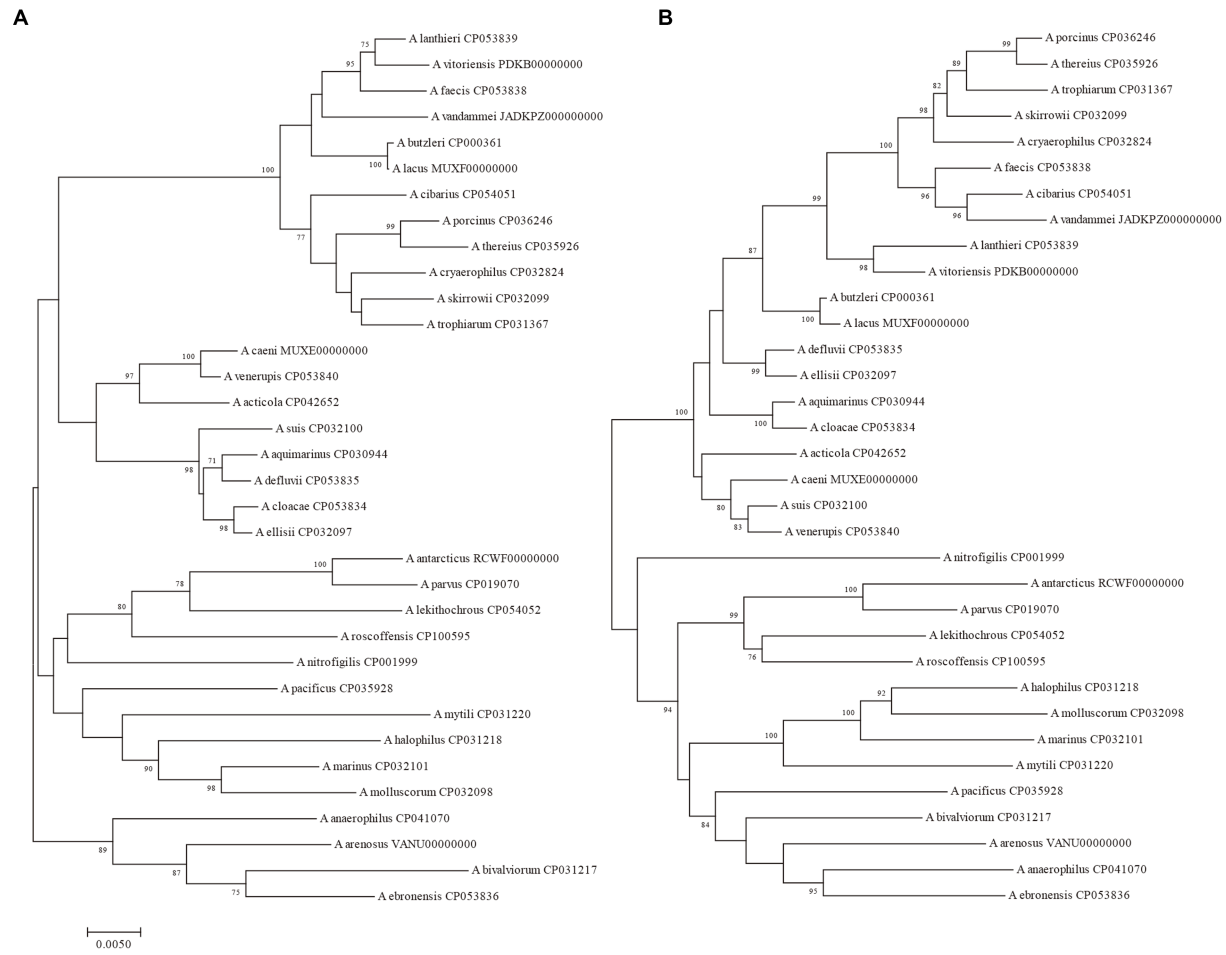


FIGURE 1
The NJ tree was constructed based on the 16S rRNA gene and 23S rRNA gene sequences of 34 *Arcobacter* species type strains, with a bootstrap value of 1,000. **(A)** The phylogenetic tree was constructed using the 16S rRNA gene, and **(B)** was constructed using the 23S rRNA gene. Bar indicated 5 substitutions per 1,000 bp.

differentiate between *A. butzleri* and *A. lacus*. For most species of *Arcobacter*, phylogenetic trees based on 16S rRNA and 23S rRNA genes have better resolution.

Species classification and genetic population

The results of the ANI and the *isDDH* calculations among the studied genomes were given in Table 2, Supplementary Table S3, and Figure 3. Significant differences in ANI were observed among different species of *Arcobacter*. The ANI values among some strains within *A. cloacae*, *A. lanthieri*, *A. marinus*, *A. skirrowii*, and *A. cryaerophilus* species were < 96%, and the *isDDH* values were < 70%. Among them, the most significant differences in ANI and *isDDH* were observed between subspecies of *A. cryaerophilus*, with ANI and *isDDH* values of 92.32 and 48.10%, respectively. However, the ANI or *isDDH* values within the species were significantly higher than those with the closest related species. In addition to the known species of *Arcobacter*, 17 genomes potentially represented 14 new species that were identified. The ANI values between these new species and the known genomes of *Arcobacter* exhibited significant differences. The ANI and *isDDH* values compared to known *Arcobacter* species were below 96 and 70%, respectively, which were the cut-off values proposed for delineating new species. Only the ANI between *A.* spp._PDJV00000000 and *A. nitrofigilis*_CP001999 > 90%, while for the remaining genomes <90%.

Phylogenetic reconstruction using the marker gene

The analysis of 371 genomes revealed a total of 835,009 genes 10,652 orthogroups, and 3,395 unassigned genes. Among these orthogroups, 216 were found to be present in all analyzed genomes, with 84 of them being single-copy orthologous genes. To elucidate the taxonomic relationships among members of the *Arcobacter* genus, we constructed a high-quality NJ phylogenomic tree based on the concatenation of these 84 conserved single-copy orthologous genes (Figure 4). The phylogenetic tree, derived from 84 single-copy homologous genes, demonstrated excellent resolution in identifying *Arcobacter* species. Notably, even *A. butzleri* and *A. lacus*,

10.3389/fmicb.2023.1278268

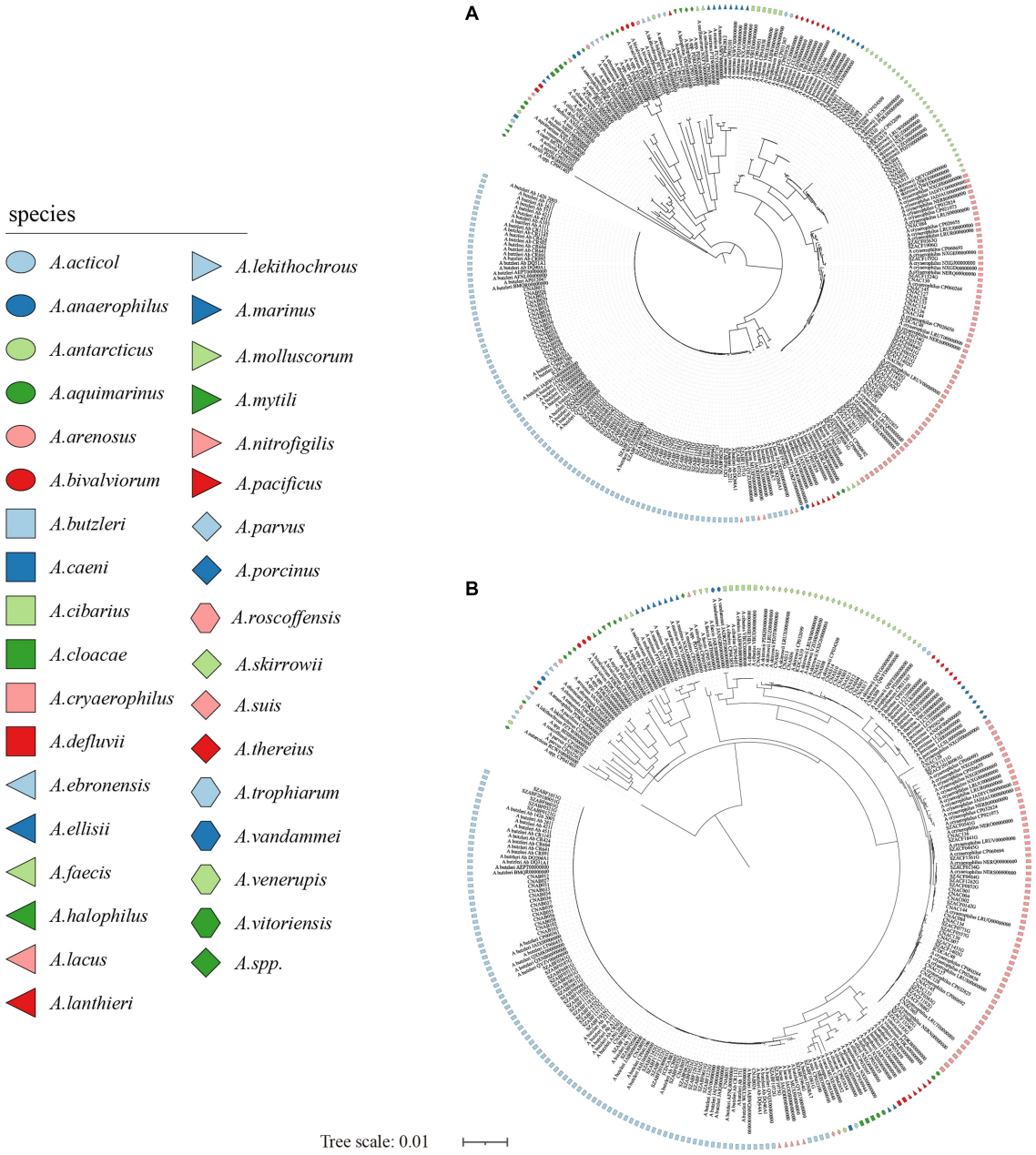


FIGURE 2
The NJ tree was constructed based on the 16S rRNA gene and 23S rRNA gene sequences of 281 *Arcobacter* genomes, with a bootstrap value of 1,000. **(A)** The phylogenetic tree was constructed using the 16S rRNA gene, and **(B)** was constructed using the 23S rRNA gene. Different colors or shapes indicated different *Arcobacter* species. Bar indicated 1 substitution per 100 bp.

characterized by remarkably high ANI values, can be clearly differentiated. Remarkably, the species classification results derived from the phylogenetic tree using the 84 single-copy homologous genes closely aligned with the ANI results, which meant that *Arcobacter* can be accurately classified using single-copy concatenation genes. Phylogenetic trees for each single-copy orthologous gene were also constructed using nucleotide and amino acid sequences. When comparing the phylogenetic trees constructed based on nucleotide and amino acid sequences of each gene with ANI results, it was found that the topology of the phylogenetic tree built using gene711 was nearly identical to the phylogenetic tree constructed using the concatenation of 84 single-copy homologous genes (Figure 5; Supplementary Figure S2). During the sequence alignment analysis of each gene, gene711 effectively differentiated all species within the *Arcobacter* genus. Furthermore, the sequence similarities within species were found to be >96% (except for *A. cryaerophilus* and *A. marinus*), while the maximum sequence similarity between different species was <94%. Consequently, gene711 could be considered a reliable signature gene for identifying *Arcobacter* species, with a sequence similarity threshold of greater than 95–96% defining the same species (Table 2; Supplementary Tables S2, S3).

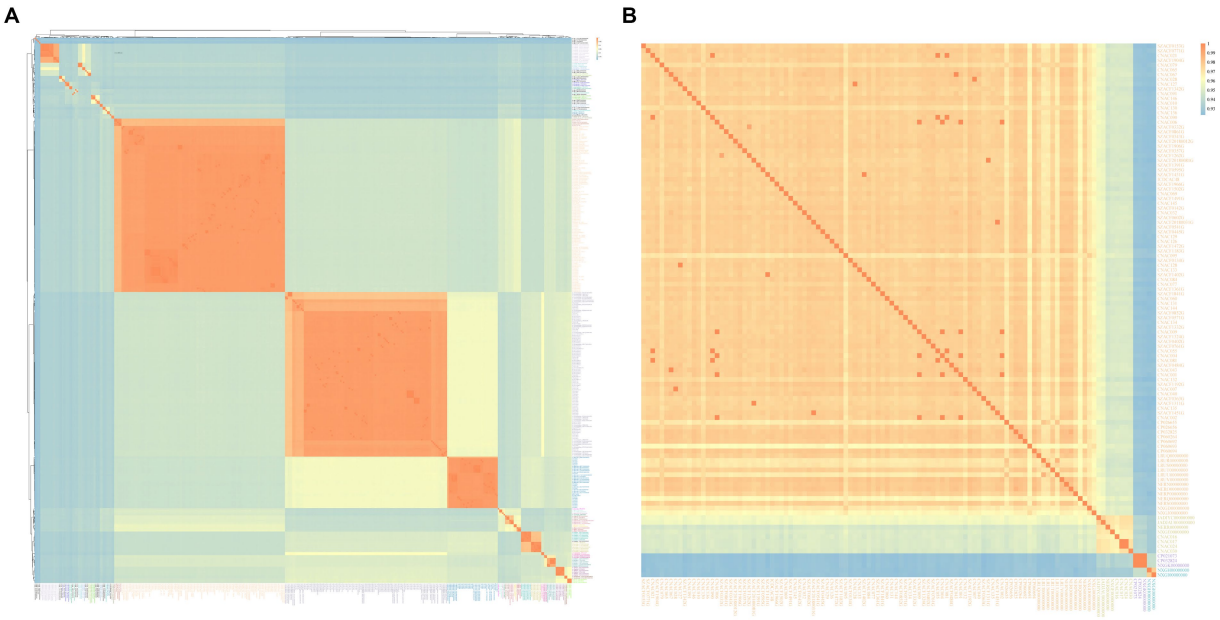


FIGURE 3
Arcobacter ANIb heatmap using the pheatmap package. **(A)** was the ANIb heatmap of 34 known *Arcobacter* species and 14 unknown *Arcobacter* species, and **(B)** was the heatmap of *A. cryaerophilus*. The depth of the color indicated the size of the ANI value, which increased sequentially from blue to orange.

Arcobacter cloacae, *Arcobacter lanthieri*, *Arcobacter skirrowii*, *Arcobacter marinus*, and *Arcobacter cryaerophilus* classification using the marker gene

The gene711 exhibited sequence similarity above 96% in *A. cloacae*, *A. lanthieri*, and *A. skirrowii*, while within these species, their ANI and *isDDH* values were below the classification thresholds of 96 and 70%, respectively. In *A. marinus*, *A.marinus*_CP042812, *A. marinus*_NWVW00000000, and *A. marinus*_PTIW00000000 showed gene711 sequence similarities ranging between 95 and 96% with other genomes, which was consistent with the ANI and *isDDH* results. For *A. cryaerophilus*, except for CNAC091 and *A. cryaerophilus*_NERP00000000, gene711 effectively divided *A. cryaerophilus* into four distinct subspecies, as shown in Figures 3B, 6 and Supplementary Table S3. The sequence similarity of gene711 was >96% within each subspecies, while the sequence similarity between subspecies was <96%, similar to the results based on ANI and *isDDH*.

Discussion

Arcobacter is recognized as a globally emerging foodborne and zoonotic pathogen with a wide range of sources and regions (Collado and Figueras, 2011; Ferreira et al., 2016). Understanding its genomic and classification characteristics is crucial for further investigations of this pathogen. In this study, a total of 371 genomes, comprising 34 named *Arcobacter* species and 14 unclassified *Arcobacter* species, were selected to elucidate the taxonomic characteristics of *Arcobacter*. The quality of the genome sequences generally met the minimal standards established for using genome data for taxonomical purposes (Chun et al., 2018). Globally, the genome size ranged from 1.68 Mb to 3.57 Mb. The G + C content ranged from 26.08 to 31.00%. Significant variations in genome size and GC content were observed in *Arcobacter*, suggesting considerable genomic diversity and divergence. This aspect could be one of the reasons contributing to the current challenges in the taxonomic classification of *Arcobacter*.

Like other bacterial genera, the taxonomic classification of *Arcobacter* has traditionally been based on the analysis of the 16S rRNA gene (Wesley et al., 1995). In fact, several potential new *Arcobacter* species could be inferred from the sequences available in public databases, similar to the 17 genomes downloaded in this study, which included 14 potentially new *Arcobacter* species. In previous studies, the similarity of the 16S rRNA gene has been considered a decisive characteristic for taxonomic classification at the genus or species level (Stackebrandt, 2006). Specifically, the sequence similarity of >98.7% in the 16S rRNA gene has been found to show good consistency with an *isDDH* > 70% (Stackebrandt, 2006). The sequence similarity of the 16S rRNA gene in 34 type strains of *Arcobacter* among multiple species was observed to be >98.7%. Moreover, expanding the number of 16S rRNA gene sequences to 281 revealed that more species displayed 16S rRNA gene sequence similarities >98.7%. However, it was necessary to note that phylogenetic trees constructed solely based on the 16S rRNA gene could cluster individuals of the same species together; however, relying solely on the 98.7% similarity threshold for species classification might lead to biased results. In other words, the discriminatory power of the 16S rRNA gene was limited when dealing with species that possessed highly similar 16S rRNA gene sequences. The 23S rRNA gene sequences were also attempted to assess *Arcobacter* interspecies differences, as published data indicated 16S rRNA gene sequences did not contain sufficient information to effectively discriminate between strains (Deshpande et al., 2013). However, our findings indicated that the 23S rRNA gene sequences were also insufficient for effective discrimination, likely due

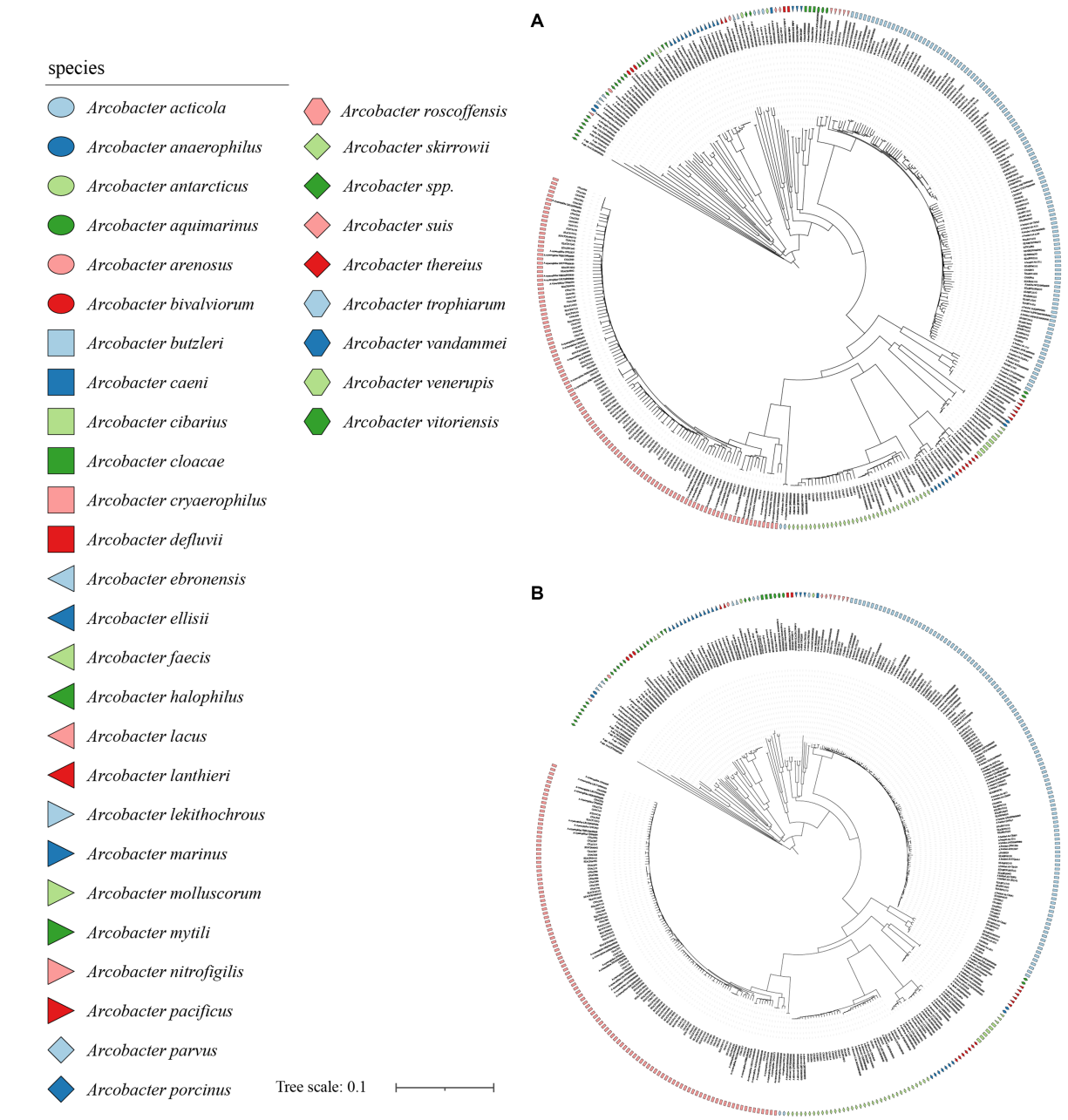


FIGURE 4
The NJ tree was constructed based on the 84 single-copy homologous genes, with a bootstrap value of 1,000. **(A)** was the phylogenetic tree constructed using nucleotide sequence, and **(B)** was the phylogenetic tree constructed using amino acid sequence. Different colors or shapes indicated different *Arcobacter* species. Bar indicated 1 substitution per 10 bp.

to the increased burden of additional sequences. Despite our efforts, the results obtained using the 23S rRNA gene were similar to those obtained using the 16S rRNA gene, further underscoring the limited discriminatory power for species with high sequence similarity.

Nowadays, genomic data such as the ANI and the *isDDH* are being increasingly used to define bacterial species, although their full potential for delineating genera has yet to be explored (Perez-Cataluna et al., 2018b; On et al., 2021). As discussed in other studies, the ANI and *isDDH* indices have been proven to provide reliable information for the delineation of *Arcobacter* species and have also been included in the minimal guidelines for defining species using genomes (Chun et al., 2018; Perez-Cataluna et al., 2018b; On et al., 2021). For *Arcobacter*, ANI values >96% were the ones that better correlated with *isDDH* results >70% in previous studies (Perez-Cataluna et al., 2018a; Zhou et al., 2022), which was further confirmed in this study. The ANI values between genomes of most *Arcobacter* species were consistent at >96%, except for certain genomes in *A. cloacae*, *A. lanthieri*, *A. marinus*, *A. skirrowii*, and *A. cryaerophilus* that did not meet the 96% classification threshold. Additionally, *isDDH* analysis was performed on species with ANI values <96%, and the results were consistent with the ANI result. Specifically, for genomes with ANI values<96%, their *isDDH* values

10.3389/fmicb.2023.1278268

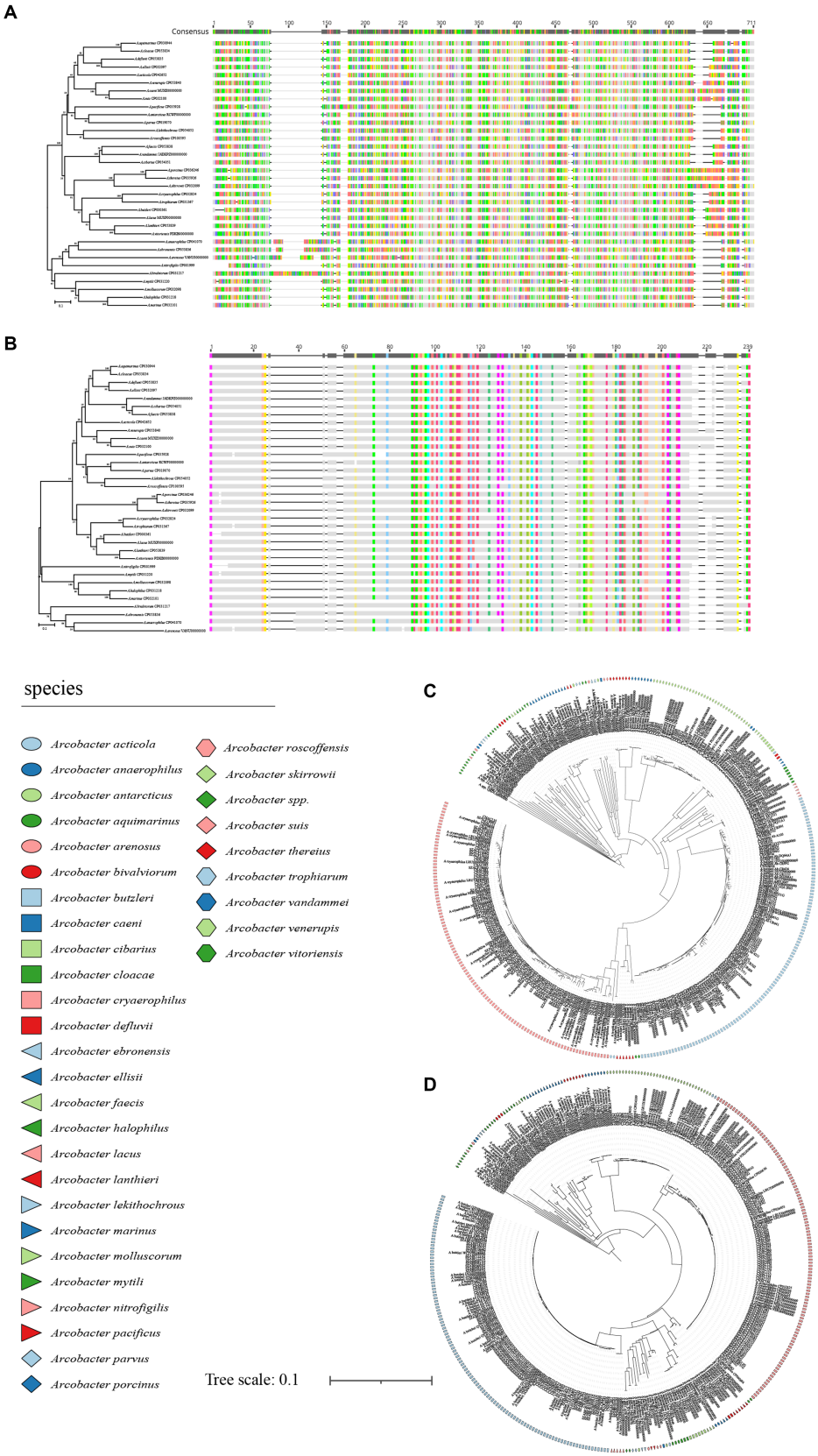


FIGURE 5
The NJ tree was constructed based on gene711, with a bootstrap value of 1,000. **(A)** was the phylogenetic tree constructed using the nucleotide sequence of 34 *Arcobacter* species type strain, **(B)** was the phylogenetic tree constructed using amino acid sequence of 34 *Arcobacter* species type strain, **(C)** was the phylogenetic tree constructed using the nucleotide sequence of 371 *Arcobacter* genomes, **(D)** was the phylogenetic tree constructed using amino acid sequence of 371 *Arcobacter* genomes. Different colors or shapes indicated different *Arcobacter* species.

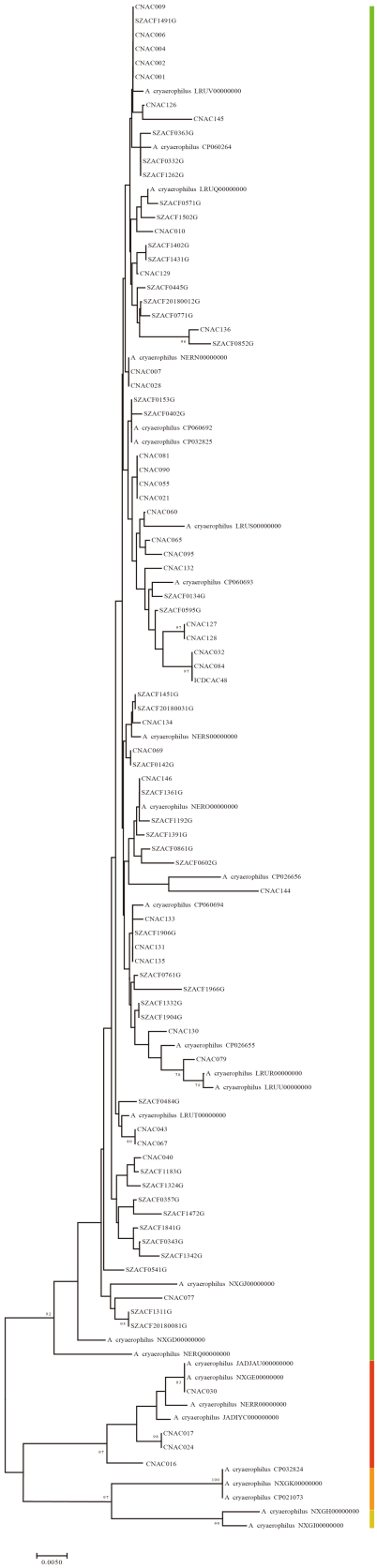


FIGURE 6
The phylogenetic tree was generated based on the sequences of gene711. The neighbor-joining method was used to generate the phylogenetic tree, which was performed using MEGA 7.0 with 1,000 bootstrap replications. Bars of different colors represented different subclades. Bar indicated 5 substitutions per 1,000 bp.

were found to be <70%. For ANIm, intraspecies pairs generally have >96% identity, while interspecies pairs generally have <93%, with an intermediate range of 93–96% where species circumspection cannot be assured (Rossello-Mora and Amann, 2015). These findings suggested substantial genomic differences within *Arcobacter* species, even though they could be classified into different subspecies. Previous studies have proposed that *A. cryaerophilus* should be divided into four subspecies according to the species classification criteria of ANI values >96% and *isDDH* values >70% (Zhou et al., 2022), which was further confirmed in this study. Within the *Arcobacter* genus, 17 genomes potentially represented 14 new potentially species. The ANI values between these new species and the known genomes of *Arcobacter* exhibited significant differences. Only the ANI between *A.* spp._PDJV00000000 and *A. nitrofigilis*_CP001999 > 90% and reached 91.64%, while the ANI for the remaining genomes <90%. These findings further emphasized the substantial genomic diversity within the *Arcobacter* genus, which posed challenges for population classification.

This study established a method based on the construction of phylogenetic trees using single-copy orthologous genes for the rapid and simplified classification of *Arcobacter* species. A robust means of species identification within *Arcobacter* was provided by utilizing 84 single-copy orthologous genes. However, this method was not widely endorsed due to its reliance on a considerable number of genes. Fortunately, we have discovered that gene711 effectively differentiated various species within *Arcobacter*. The gene711, which encoded a 186–218 amino acid in *Arcobacter*, was a FlgO family outer membrane protein and was capable of reproducing a tree with a similar topology to our genome-based phylogeny. The gene711 sequences demonstrated high nucleotide diversity and yielded a tree that accurately separates strains into phylogenetic groups defined by ANI-based analysis. The gene711 exhibited sequence similarity >96% within the same species, while the similarity between different species was significantly <96%. The neighboring genes upstream and downstream of gene711 also displayed relatively conserved characteristics, making them potential targets for developing sequence-based analysis or real-time PCR assays to detect *Arcobacter* species. The discriminatory power of the gene711 locus made it possible to improve the accuracy of species identification within the *Arcobacter* genus. As mentioned earlier, certain genomes within *A. cloacae*, *A. lanthieri*, *A. marinus*, *A. skirrowii*, and *A. cryaerophilus* did not meet the species classification criteria of ANI values >96% and *isDDH* values >70% within the same species. Among these species, we used gene711 to verify and found that except for *A. marinus* and *A. cryaerophilus*, the remaining species met the requirement of gene711 > 96% within the species and gene711 < 96% between species. Previous studies (Zhou et al., 2022) have identified four subspecies within *A. cryaerophilus*, and our study using gene711 for *A. cryaerophilus* subspecies classification further supported this conclusion. However, there were also instances of gene711 anomalies in certain strains within *A. cryaerophilus*, such as CNAC091.

To our knowledge, this is the first time that gene711 has been used as a phylogenetic marker within a bacterial genus. As highlighted in the review by Collado and Figueras (2011), numerous uncultured or as-yet-undescribed species of *Arcobacter* have been identified based on nearly full-length 16S rRNA gene sequences, potentially surpassing the number of already known species at that time. The emergence of new species can be anticipated in the near future, further validating the significance of gene711 proposed in this study.

Conclusion

In this study, we evaluated the efficacy of various genome-based phylogenetic tools in discriminating between different *Arcobacter* species. Novel approaches for the classification of the *Arcobacter* were employed in this study. Finally, a maker gene (gene711) that demonstrated greater discriminatory power and robustness than other commonly used markers was identified, making it a valuable tool for future molecular identification of *Arcobacter* species. In summary, our study offers valuable insights into the evolution, genetic diversity, and species classification of *Arcobacter*, thereby shedding new light on the behavior and characteristics of this genus.

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

GZ: Methodology, Software, Writing – original draft, Writing – review & editing. YG: Writing – review & editing. HW: Software, Writing – review & editing. XC: Writing – review & editing. XZ: Writing – review & editing. ZS: Supervision, Writing – review & editing. XY: Supervision, Writing – review & editing. JZ: Supervision, Writing – review & editing. MZ: Methodology, 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 Key Research and Development Program of China under Grant 2021YFC2301000; the Project for Novel Detection Techniques of Bacterial Pathogens under Grant 32073; Prevention and Intervention of Bacterial and Fungal Infectious Diseases under Grant 102393220020020000031; and Enhancement of Comprehensive Monitoring, Prevention, and Control Capabilities for Traditional Infectious Diseases Such as Plague, Cholera, and Brucellosis under Grant 102393230020020000002.

Acknowledgments

We thank our colleagues from Nanshan Center for Disease Control and Prevention, Shunyi District Center for Disease Control and Prevention, and the Chinese Center for Disease Control and Prevention.

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.2023.1278268/full#supplementary-material

References

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., et al. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 68, 461–466. doi: 10.1099/ijsem.0.002516

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862. doi: 10.1128/CMR.17.4.840-862.2004

Collado, L., and Figueras, M. J. (2011). Taxonomy, epidemiology, and clinical relevance of the genus *Arcobacter*. *Clin. Microbiol. Rev.* 24, 174–192. doi: 10.1128/CMR.00034-10

Debruyne, L., Houf, K., Douidah, L., De Smet, S., and Vandamme, P. (2010). Reassessment of the taxonomy of *Arcobacter cryaerophilus*. *Syst. Appl. Microbiol.* 33, 7–14. doi: 10.1016/j.syapm.2009.10.001

Deshpande, N. P., Kaakoush, N. O., Wilkins, M. R., and Mitchell, H. M. (2013). Comparative genomics of *Campylobacter concisus* isolates reveals genetic diversity and provides insights into disease association. *BMC Genomics* 14:585. doi: 10.1186/1471-2164-14-585

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

Ferreira, S., Queiroz, J. A., Oleastro, M., and Domingues, F. C. (2016). Insights in the pathogenesis and resistance of *Arcobacter*: a review. *Crit. Rev. Microbiol.* 42, 364–383. doi: 10.3109/1040841X.2014.954523

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

Hanel, I., Tomaso, H., and Neubauer, H. (2016). *Arcobacter* - an underestimated zoonotic pathogen? *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 59, 789–794. doi: 10.1007/s00103-016-2350-7

Ho, H. T., Lipman, L. J., and Gaastra, W. (2006). *Arcobacter*, what is known and unknown about a potential foodborne zoonotic agent! *Vet. Microbiol.* 115, 1–13. doi: 10.1016/j.vetmic.2006.03.004

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

Isidro, J., Ferreira, S., Pinto, M., Domingues, F., Oleastro, M., Gomes, J. P., et al. (2020). Virulence and antibiotic resistance plasticity of *Arcobacter butzleri*: insights on the genomic diversity of an emerging human pathogen. *Infect. Genet. Evol.* 80:104213. doi: 10.1016/j.meegid.2020.104213

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

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi: 10.1093/nar/gkh152

Ma, Y., Ju, C., Zhou, G., Yu, M., Chen, H., He, J., et al. (2022). Genetic characteristics, antimicrobial resistance, and prevalence of *Arcobacter* spp. isolated from various sources in Shenzhen, China. *Front. Microbiol.* 13:1004224. doi: 10.3389/fmicb.2022.1004224

On, S. L. (2001). Taxonomy of *Campylobacter*, *Arcobacter*, *Helicobacter* and related bacteria: current status, future prospects and immediate concerns. *Symp Ser Soc Appl Microbiol* 30, 1S–15S. doi: 10.1046/j.1365-2672.2001.01349.x

On, S. L. W., Miller, W. G., Biggs, P. J., Cornelius, A. J., and Vandamme, P. (2020). A critical rebuttal of the proposed division of the genus *Arcobacter* into six genera using comparative genomic, phylogenetic, and phenotypic criteria. *Syst. Appl. Microbiol.* 43:126108. doi: 10.1016/j.syapm.2020.126108

On, S. L. W., Miller, W. G., Biggs, P. J., Cornelius, A. J., and Vandamme, P. (2021). *Aliarcobacter*, *Halarcobacter*, *Malaciobacter*, *Pseudarcobacter* and *Poseidonibacter* are later synonyms of *Arcobacter*: transfer of *Poseidonibacter parvus*, *Poseidonibacter antarcticus*, '*Halarcobacter arenosus*', and '*Aliarcobacter vitoriensis*' to *Arcobacter* as *Arcobacter parvus* comb. nov., *Arcobacter antarcticus* comb. nov., *Arcobacter arenosus* comb. nov. and *Arcobacter vitoriensis* comb. nov. *Int. J. Syst. Evol. Microbiol.* 71:5133. doi: 10.1099/ijsem.0.005133

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

Pascual, J., Lepleux, C., Methner, A., Sproer, C., Bunk, B., and Overmann, J. (2023). *Arcobacter roscoffensis* sp. nov., a marine bacterium isolated from coastal seawater. *Int. J. Syst. Evol. Microbiol.* 73:5895. doi: 10.1099/ijsem.0.005895

Perez-Cataluna, A., Collado, L., Salgado, O., Lefinanco, V., and Figueras, M. J. (2018a). A polyphasic and taxogenomic evaluation uncovers *Arcobacter cryaerophilus* as a species complex that embraces four Genomovars. *Front. Microbiol.* 9:805. doi: 10.3389/fmicb.2018.00805

Perez-Cataluna, A., Salas-Masso, N., Dieguez, A. L., Balboa, S., Lema, A., Romalde, J. L., et al. (2018b). Revisiting the taxonomy of the genus *Arcobacter*: getting order from the Chaos. *Front. Microbiol.* 9:2077. doi: 10.3389/fmicb.2018.02077

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rathlavath, S., Kumar, S., and Nayak, B. B. (2017). Comparative isolation and genetic diversity of *Arcobacter* sp. from fish and the coastal environment. *Lett. Appl. Microbiol.* 65, 42–49. doi: 10.1111/lam.12743

Rossello-Mora, R., and Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* 38, 209–216. doi: 10.1016/j.syapm.2015.02.001

Roth, A., Andrees, S., Kroppenstedt, R. M., Harmsen, D., and Mauch, H. (2003). Phylogeny of the genus Nocardia based on reassessed 16S rRNA gene sequences reveals underspeciation and division of strains classified as *Nocardia asteroides* into three established species and two unnamed taxons. *J. Clin. Microbiol.* 41, 851–856. doi: 10.1128/JCM.41.2.851-856.2003

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

Stackebrandt, E. E. J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microb. Today* 33:152.

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

Van den Abeele, A. M., Vogelaers, D., Van Hende, J., and Houf, K. (2014). Prevalence of *Arcobacter* species among humans, Belgium, 2008-2013. *Emerg. Infect. Dis.* 20, 1731–1734. doi: 10.3201/eid2010.140433

Van Driessche, E., Houf, K., Vangroenweghe, F., De Zutter, L., and Van Hoof, J. (2005). Prevalence, enumeration and strain variation of *Arcobacter* species in the faeces of healthy cattle in Belgium. *Vet. Microbiol.* 105, 149–154. doi: 10.1016/j.vetmic.2004.11.002

Vandamme, P., Falsen, E., Rossau, R., Hoste, B., Segers, P., Tytgat, R., et al. (1991). Revision of *Campylobacter*, *Helicobacter*, and *Wolinella* taxonomy: emendation of generic descriptions and proposal of *Arcobacter* gen. Nov. *Int. J. Syst. Bacteriol.* 41, 88–103. doi: 10.1099/00207713-41-1-88

Vandamme, P., Vancanneyt, M., Pot, B., Mels, L., Hoste, B., Dewettinck, D., et al. (1992). Polyphasic taxonomic study of the emended genus *Arcobacter* with *Arcobacter butzleri* comb. nov. and *Arcobacter skirrowii* sp. nov., an aerotolerant bacterium isolated from veterinary specimens. *Int. J. Syst. Bacteriol.* 42, 344–356. doi: 10.1099/00207713-42-3-344

Vandenberg, O., Dediste, A., Houf, K., Ibekwem, S., Souayah, H., Cadranel, S., et al. (2004). *Arcobacter* species in humans. *Emerg. Infect. Dis.* 10, 1863–1867. doi: 10.3201/eid1010.040241

Wang, Y. Y., Zhou, G. L., Li, Y., Gu, Y. X., He, M., Zhang, S., et al. (2021). Genetic characteristics and antimicrobial susceptibility of *Arcobacter butzleri* isolates from raw chicken meat and patients with diarrhea in China. *Biomed. Environ. Sci.* 34, 1024–1028. doi: 10.3967/bes2021.139

Wesley, I. V., Schroeder-Tucker, L., Baetz, A. L., Dewhirst, F. E., and Paster, B. J. (1995). *Arcobacter*-specific and *Arcobacter butzleri*-specific 16S rRNA-based DNA probes. *J. Clin. Microbiol.* 33, 1691–1698. doi: 10.1128/jcm.33.7.1691-1698.1995

Whitman, W. B. (2015). Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst. Appl. Microbiol.* 38, 217–222. doi: 10.1016/j.syapm.2015.02.003

Zhou, G., Wang, M., Wang, H., Chen, X., Gu, Y., Shao, Z., et al. (2022). Species classification and novel plasmid identifications in *Arcobacter cryaerophilus* and *Arcobacter cryaerophilus*-like organisms. *Front. Microbiol.* 13:984450. doi: 10.3389/fmicb.2022.984450

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 October 2023
DOI 10.3389/fmicb.2023.1275616

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Sadiya Parveen,
Johns Hopkins University, United States
Leny Jose,
Indiana University–Purdue University Indianapolis, United States

*CORRESPONDENCE
Bum-Joon Kim
✉ kbumjoon@snu.ac.kr

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

RECEIVED 10 August 2023
ACCEPTED 29 September 2023
PUBLISHED 13 October 2023

CITATION
Lee D, Kim DH, Seo H, Choi S and Kim B-J (2023) Phylogenetic distribution of malonate semialdehyde decarboxylase (MSAD) genes among strains within the genus *Mycobacterium*: evidence of MSAD gene loss in the evolution of pathogenic mycobacteria. *Front. Microbiol.* 14:1275616. doi: 10.3389/fmicb.2023.1275616

COPYRIGHT
© 2023 Lee, Kim, Seo, Choi and Kim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Phylogenetic distribution of malonate semialdehyde decarboxylase (MSAD) genes among strains within the genus *Mycobacterium*: evidence of MSAD gene loss in the evolution of pathogenic mycobacteria

Duhyung Lee[1,2†], Dong Hyun Kim[1,3†], Hyejun Seo[1,2,4], Seaone Choi[1,3,5] and Bum-Joon Kim[1,2,3,4,5,6*]

[1]Department of Microbiology and Immunology, College of Medicine, Seoul National University, Seoul, Republic of Korea, [2]Cancer Research Institute, College of Medicine, Seoul National University, Seoul, Republic of Korea, [3]Department of Biomedical Sciences, College of Medicine, Seoul National University, Seoul, Republic of Korea, [4]Seoul National University Medical Research Center, Seoul, Republic of Korea, [5]Brain Korea 21 FOUR Biomedical Science Project, Seoul National University College of Medicine, Seoul, Republic of Korea, [6]Liver Research Institute, College of Medicine, Seoul National University, Seoul, Republic of Korea

Despite the great diversity of malonate semialdehyde decarboxylases (MSADs), one of five subgroups of the tautomerase superfamily (TSF) found throughout the biosphere, their distribution among strains within the genus *Mycobacterium* remains unknown. In this study, we sought to investigate the phylogenetic distribution of MSAD genes of mycobacterial species via genome analysis of 192 different reference *Mycobacterium* species or subspecies retrieved from NCBI databases. We found that in a total of 87 of 192 strains (45.3%), MSAD-1 and MSAD-2 were distributed in an exclusive manner among *Mycobacterium* species except for 12 strains, including *Mycobacterium chelonae* members, with both in their genome. Of note, *Mycobacterium* strains better adapted to the host and of high virulence potential, such as the *Mycobacterium tuberculosis* complex, *Mycobacterium leprae*, *Mycobacterium marinum*, *Mycobacterium ulcerans*, and *Mycobacterium avium* subsp. *paratuberculosis*, had no orthologs of MSAD in their genome, suggesting MSAD loss during species differentiation in pathogenic slow-growing *Mycobacterium*. To investigate the MSAD distribution among strains of *M. avium* subspecies, the genome sequences of a total of 255 reference strains from the four subspecies of *M. avium* (43 of subspecies *avium*, 162 of subspecies *hominissuis*, 49 of subspecies *paratuberculosis*, and 1 of subspecies *silvaticum*) were further analyzed. We found that only 121 of 255 strains (47.4%) had MSADs in their genome, with none of the 49 *M. avium* subsp. *paratuberculosis* strains having MSAD genes. Even in 13 of 121 *M. avium* strains with the MSAD-1 gene in their genome, deletion mutations in the 98th codon causing premature termination of MSAD were found, further highlighting the occurrence of MSAD pseudogenization during species or subspecies differentiation of *M. avium*. In conclusion, our data indicated that there are two distinct types of MSADs, MSAD-1 and MSAD-2, among strains in the *Mycobacterium* genus, but more than half of the strains, including pathogenic mycobacteria, *M. tuberculosis* and *M. leprae*,

have no orthologs in their genome, suggesting MSAD loss during host adaptation of pathogenic mycobacteria. In the future, the role of two distinct MSADs, MSAD-1 and MSAD-2, in mycobacterial pathogenesis or evolution should be investigated.

KEYWORDS

malonate semialdehyde decarboxylase (MSAD), *Mycobacterium*, phylogenetic distribution, MSAD-1, MSAD-2, *Mycobacterium avium* subsp. *paratuberculosis*

1. Introduction

The genus *Mycobacterium* is composed of more than 200 established and validated species and subspecies belonging to the phylum *Actinobacteria*, defined by a rod-shaped morphology, acid fastness, unusual cell walls containing mycolic acids, and relatively high genomic DNA G+C contents (∼ 61 to 71%) (Vincent Lévy-Frébault and Portaels, 1992; Jankute et al., 2015; Gupta et al., 2018). Generally, *Mycobacterium* can be separated into two groups according to their pathogenic potential: human pathogens, including *Mycobacterium tuberculosis* and *Mycobacterium leprae*, which cause tuberculosis and leprosy, respectively, and nontuberculous *Mycobacterium* (NTM), which are environmental mycobacteria that do not cause tuberculosis, as the name suggests, and are often nonpathogenic to humans and animals (Gagneux, 2018; Johansen et al., 2020; van Hooij and Geluk, 2021). The genus can be further separated into two groups, slow-growing *Mycobacterium* (SGM) (i.e., requiring more than 7 days to form visible colonies on solid agar) and rapid-growing *Mycobacterium* (RGM) requiring <7 days to form colonies (Lotti and Hautmann, 1993). Although most NTMs are found in the environment, such as soil or natural and drinking water sources, a few species, including the *Mycobacterium avium complex* (MAC) and *Mycobacterium abscessus*, often cause serious lung diseases through infections in humans (Nishiuchi et al., 2017; Johansen et al., 2020; Ratnatunga et al., 2020). The increases in immunosuppressive drug use, broad-spectrum antibiotic therapy, and patients with underlying lung diseases, including cystic fibrosis and bronchiectasis, have contributed to the recent rise in the global incidence of NTM infections in developed countries (Petrini, 2006; Chalmers et al., 2018; To et al., 2020).

The tautomerase superfamily (TSF) consists of more than 11,000 members throughout the biosphere, which can be classified into five major categories: 4-oxalocrotonate tautomerase (4-OT), 5-(carboxymethyl)-2-hydroxymuconate isomerase (CHMI), macrophage migration inhibitory factor (MIF), *cis*- and *trans*-3-chloroacrylic acid dehalogenase (*cis*-CaaD and CaaD, respectively), and malonate semialdehyde decarboxylase (MSAD) (Poelarends et al., 2008a; Davidson et al., 2018). Despite some exceptions, they share a β-α-β structure with an unusual catalytic amino-terminal proline as a general base. MSAD not only catalyzes the decarboxylation of malonate semialdehyde to produce acetaldehyde but also participates in degradative pathways of 1,3-dichloropropene, a soil fumigant (Poelarends and Whitman, 2004).

To date, structural studies regarding MSAD have been mainly focused on nine organisms, originating from various species such as *Pseudomonas pavonaceae* 170 (PpMSAD) (Poelarends et al., 2003, 2004, 2005; Almrud et al., 2005), *Coryneform bacterium* strain fg41 (FG41 MSAD) (Poelarends et al., 2008b; Guo et al., 2013), *Lactobacillus casei* strain BL23 (IolK) (Poelarends et al., 2008b), *Bacillus subtilis* strain 168 (YusQ, YodA, and YrdN) (Poelarends et al., 2005), *Burkholderia phymatum* strain STM815 (Bp4401) (Huddleston et al., 2014), *Calothrix* sp. PCC 6303(437) (Lancaster et al., 2022), and *Rivularia* sp. PCC 7116(JJ3) (Lancaster et al., 2022). However, research on the MSADs of *Mycobacterium* strains is limited.

Comparative genomic studies have revealed that overt human-pathogenic *Mycobacterium* species, including *M. tuberculosis, M. leprae*, and *Mycobacterium ulcerans,* have undergone genome reduction and gene loss since their evolution from the ancestor (Gutierrez et al., 2005; Marri et al., 2006; Gómez-Valero et al., 2007; Qi et al., 2009). Since MSAD plays a key role in bacterial metabolism, it may have had distinct effects on the evolutionary scenario of *Mycobacterium* species in terms of their groups, pathogenic or environmental strains and slow-growing or rapid-growing status. Therefore, investigation of the phylogenetic distribution of MSAD genes among *Mycobacterium* strains would provide novel insight into their evolution and pathogenesis.

In the present study, we sought to investigate the phylogenetic distribution of MSAD genes of mycobacterial species via genome analysis of 192 different reference *Mycobacterium* species or subspecies retrieved from NCBI databases. In addition, to further test our hypothesis of MSAD loss in *Mycobacterium* strains more adapted to host-associated life, we further checked the distribution of MSADs among 255 MAC strains of the four subspecies of *M. avium* retrieved from NCBI databases.

2. Materials and methods

2.1. *Mycobacterium* type strain database

A total of 192 type strains of *Mycobacterium* with whole genome sequences present in the NCBI taxonomy browser, ATCC (American Type Culture Collection), DSMZ (German Collection of Microorganisms and Cell Cultures), and JCM (Japan Collection of Microorganisms) were selected (**Figure 1**). Of these, 129 type strains were referenced from a previous study (Gupta et al., 2018), and the other 67 type strains were selected from the ATCC (21 strains), DSMZ (34 strains), and JCM (4 strains). In addition, we selected four newly identified *Mycobacterium* species, *Mycobacterium dioxanotrophicus*, *Mycolicibacterium nivoides*, *Mycobacterium terramassiliense*, and *Mycobacterium senriense* (He

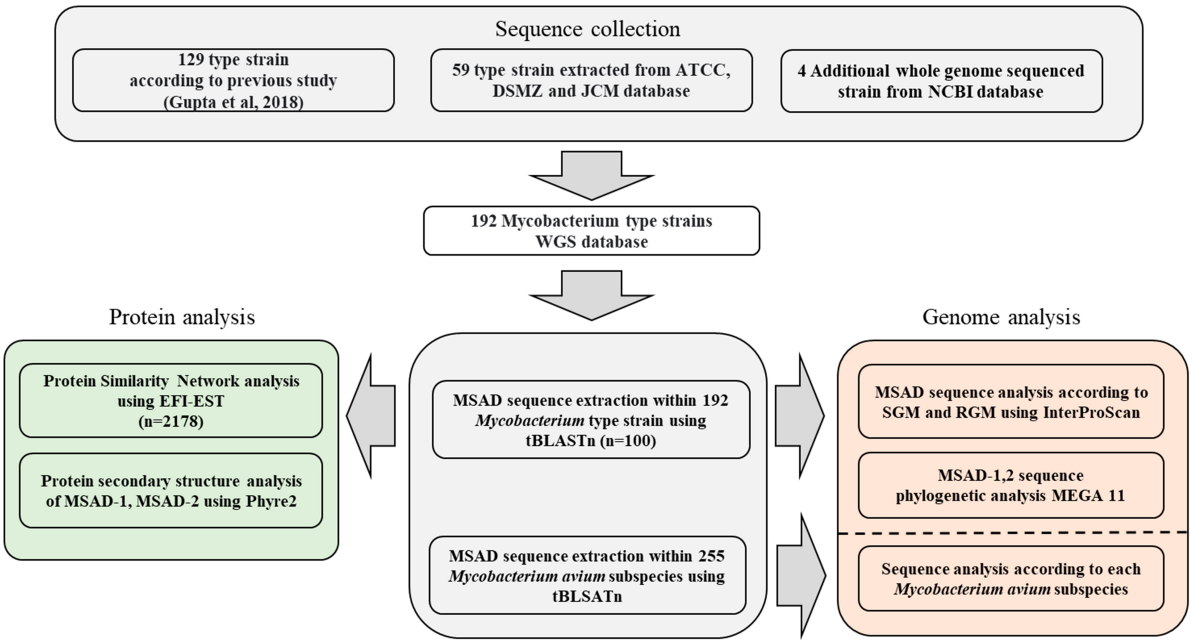


FIGURE 1
Workflow representing the construction of the *Mycobacterium* genome sequence database and software used for each analysis in this study.

et al., 2017; Bouam et al., 2018; Dahl et al., 2021; Abe et al., 2022), with their whole genome sequences registered with the NCBI and added them to the database. The whole-genome sequencing project names and accession numbers are listed in **Supplementary Table 1**.

2.2. MSAD sequence data collection

Using two MSAD sequences (MSAD-1: NCBI accession number OLT81431.1 and MSAD-2: OLT83119.1) of *Mycobacterium chelonae* subsp. *chelonae* ATCC 35752^{T} as query sequences, the presence of MSAD proteins was confirmed in 192 *Mycobacterium* type strains through the tBLASTn algorithm (**Supplementary Table 1**; Altschul et al., 1990, 1997). Among the BLAST search results, sequences were extracted based on a cutoff of 90% for query coverage and 10^{-20} for the *E*-value. MSAD sequences ranging in length from 119 to 137 amino acids were collected. A total of 100 MSAD sequences from 192 *Mycobacterium* type strains were extracted. The extracted protein sequences of MSAD orthologs are stored in FASTA format (**Supplementary Data 1**) and their DNA sequences are listed in **Supplementary Table 2**.

2.3. Protein similarity network of the total MSAD family

A total of 2,078 MSAD sequences, which are widely distributed throughout the biosphere, are registered in the SFLD (Superfamily of Ligand-Binding Protein Database) (Akiva et al., 2014; Davidson et al., 2018). They were trimmed and aligned before being submitted to the Enzyme Function Initiative enzyme similarity tool (EFI-EST) webserver for analysis (Zallot et al., 2019) (Job ID 96068). The EFI-EST was used to create the sequence similarity network (Oberg et al., 2023). We aligned total MSAD protein sequences (2,078 sequences) and mycobacterial MSADs (100 sequences) together and submitted them to the EFI-EST algorithm. Merged MSAD sequences were stored in FASTA format (**Supplementary Data 2**). In this study, the *E*-value for SSN Edge calculation was set to 5, and the convergence ratio was set to 0.743. This value decreases from 1.0 for sequences that are very similar (identical) to 0.0 for sequences that are very different (unrelated). Constructed networks were then transferred to the Color SSN utility for Representative Node (RepNode) Networks (Job ID 96072). The 100% identity RepNode network was stored as **Supplementary Data 3**, and 50% identical sequences were grouped to reduce the total number of nodes. The colored SSN was visualized in Cytoscape (Shannon et al., 2003) (version 3.9.1).

2.4. Protein structure prediction of MSAD

Four MSAD sequences of *M. avium* (MSAD-1 of slow growing mycobacteria), *M. abscessus* (MSAD-1 of rapid growing mycobacteria), *Mycobacterium terrae* (MSAD-2 of slow growing mycobacteria), and *Mycobacterium fortuitum* (MSAD-2 of rapid growing mycobacteria) were submitted to the phyre2 webserver[1] for protein function prediction (Kelley et al., 2015). The structures of four stains of MSAD (two strains of MSAD-1 and two strains of MSAD-2) were determined based on their protein sequences (**Supplementary Data 1**). The constructed MSAD 3D models were

1 http://www.sbg.bio.ic.ac.uk/phyre2

then compared to PDB-registered proteins. The protein derived from the *C. bacterium* strain fg41 registered in the PDB (PDB ID: 3MJZ) was used as a template to align MSAD sequence of *M. abscessus* and *M. terrae*. And the PpMSAD derived from the *P. pavonaceae* registered in the PDB (PDB ID: 2AAL) was used as a template to align MSAD sequence of *M. avium* and *M. fortuitum*. The protein structures created with phyre were analyzed with the PyMOL Molecular Graphics System Schrödinger, Inc. (version 2.5.5). The identified *Mycobacterium* type strains were used to predict the structure and function of MSAD sequences using InterProScan (Jones et al., 2014).

2.5. MSAD sequence alignment and phylogenetic analysis

A total of 57 MSAD-1 sequences and 43 MSAD-2 sequences from *Mycobacterium* reference strains were aligned using the MUSCLE method for each DNA sequence and amino acid sequence through the MEGA 11 program (Tamura et al., 2021). The DNA and protein sequence-based phylogenetic trees of MSAD-1 and MSAD-2 were constructed through the maximum-likelihood method. Branch support value was calculated through 100 bootstrap replications. Phylogenetic trees based on 644 bp *hsp65* sequences are often used to classify and identify *Mycobacterium* species (Kim et al., 2005). The *hsp65* sequence of *M. tuberculosis* H37Rv (GenBank accession number M15467) was used to extract the *hsp65* sequence from whole genome sequences of the type strains through BLAST (**Supplementary Data 4**). In this study, two *hsp65* sequence-based trees (one for MSAD-1 sequences from 56 mycobacterial strains and the other for MSAD-2 sequences from 43 strains) were constructed through MUSCLE alignment and the maximum-likelihood method. Branch support value was calculated through 100 bootstrap replications. YrdN (MSAD of *B. subtilis* strain 168) was used as an outgroup in the MSAD tree, while *Tsukamurella paurometabola* KCTC 9821^T (GenBank accession number UHIQ01000001.1) was used as an outgroup in the *hsp65* tree.

2.6. *Mycobacterium avium* subspecies strain MSAD sequence analysis

To analyze MSAD protein retention within *M. avium* subspecies, a total of 255 whole genome sequences were obtained from the NCBI database, including 43 *M. avium* subsp. *avium* sequences, 162 *M. avium* subsp. *hominissuis* sequences, 49 *M. avium* subsp. *paratuberculosis* sequences, and 1 *M. avium* subsp. *silvaticum* sequence. In total, 121 MSAD sequences were extracted from the whole genomes of *M. avium* subspecies using tBLASTn (Altschul et al., 1997). The MSAD sequences were aligned and analyzed using the MUSCLE method through MEGA 11 (Tamura et al., 2021). Information regarding the accession numbers, metadata, and all publicly available assemblies for the whole genome sequences was also extracted from the NCBI database (**Supplementary Data 5**). Extracted MSAD sequences were stored in FASTA format (**Supplementary Data 6**).

2.7. Preparation of *Mycobacterium abscessus* MSAD-1 protein

Recombinant MSAD-1 protein of *M. abscessus* (NCBI accession number OLT57519.1) were purified from *Escherichia coli* as previously described with minor modification (Jeong et al., 2022). Briefly, the DNA sequence of MSAD-1 was amplified from *M. abscessus* ATCC 19977^T using PCR with following primer sets (forward primer, 5′-TTT GGA TCC ATG CCA TTG GTG CGC ATC GAC CTC-3′; reverse primer, 5′-AAA AAG CTT GTG CGC CTG CGG CGG GCA C-3′), and cloned into pET-28a. The expression and purification of MSAD-1 were commercially commissioned by Bionics (Seoul, Republic of Korea). In detail, the protein expression was induced in *E. coli* Rosetta2 (DE3) strains (Novagen, WI, USA) transformed with pET28a-MSAD-1 by adding 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) at 26°C for 6 h. Cultured bacterial cells were harvested and sonicated for 30 cycles at 70% amplitude. After centrifuge, the supernatant was purified with HisTrap^{TM} HP His tag protein purification columns (Cytiva, MA, USA) for Ni-NTA affinity chromatography via ÄKTA go system (Cytiva, MA, USA). Purified proteins were subjected to endotoxin removal using Pierce^{TM} high-capacity endotoxin removal spin columns (Thermo Scientific, MA, USA) and quantified by Pierce^{TM} chromogenic endotoxin quant kit (Thermo Scientific, MA, USA).

2.8. Induction of pro-inflammatory cytokines, TNF-α and IL-6 on J774A.1 cells by MSAD-1

The murine macrophage cell line, J774A.1, was maintained at 37°C with 5% CO_2 in RPMI 1640 supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS). To induce protein uptake efficiently, J774A.1 cells were starved in reduced serum medium, opti-MEM (Gibco, MT, USA), at 37°C for 1 h. Subsequently, they were incubated with various concentrations of MSAD-1 or 100 ng/ml LPS (Sigma, MO, USA) in RPMI 1640 supplemented with 2% (v/v) FBS and 1% PS at 37°C for either 24 or 48 h. The culture medium was used to measure the levels of the cytokines using an enzyme-linked immunosorbent assay (ELISA) kit (Invitrogen, MA, USA) according to the manufacture's instructions.

3. Results

3.1. Construction of a novel MSAD protein similarity network including 100 mycobacterial MSAD sequences

To examine the distribution of a total of 100 mycobacterial MSAD orthologs extracted from genome sequences of 192 strains in this study among all the biospheres, we constructed a novel protein similarity network including a total of 2,178 MSAD protein sequences from established MSAD protein sequences (2,078 sequences) (Akiva et al., 2014) and mycobacterial MSADs

(100 sequences) extracted in this study and submitted to the EFI-EST algorithm (Zallot et al., 2019; **Figure 1**). All 2,178 MSADs are grouped into 12 clusters of 472 nodes, which are connected by 7,677 edges, and further clustered into two major groups (colored red and blue) (**Figure 2**). While one group (colored red) consists of 328 larger nodes of 1,633 sequences, including PpMSAD (Poelarends et al., 2003) and FG41 MSAD (Poelarends et al., 2008b), which have been widely studied for the elucidation of MSAD function, the other group consists of 81 smaller nodes of 453 MSAD sequences. We decided to designate these two groups as MSAD-1 and MSAD-2, respectively. Generally, all 100 mycobacterial MSAD sequences extracted in this study also belonged to these two major groups, MSAD-1 and MSAD-2 (**Figure 2**). However, there are some discrepancies in their distribution in the protein similarity network. Strains of MSAD-1 are more widely scattered among 13 nodes compared with those of MSAD-2, found at 4 nodes, suggesting more sequence divergence between MSAD-1 strains than between MSAD-2 strains (**Figure 2**).

3.2. The distribution of MSAD genes among *Mycobacterium* strains

Of 192 *Mycobacterium* reference strains, only fewer than half (87, 45.3%) had MSAD genes in their genomes (105 strains without any MSAD orthologs in their genomes) (**Figure 3**). The DNA sequences and accession numbers of all 100 MSADs from 87 *Mycobacterium* strains are presented in **Supplementary Table 1**. Of all 100 MSAD sequences, 57 and 43 belonged to MSAD-1 and MSAD-2, respectively. Most *Mycobacterium* strains (75 strains) have a single MSAD in their genomes, MSAD-1 or MSAD-2, in an exclusive manner (74 strains have single copy of MSAD gene and *Mycobacterium simiae* with two copies of MSAD-1). The remaining 12 species, including the three subspecies *M. chelonae* subsp. *chelonae*, *M. chelonae* subsp. *bovis* (Kim et al., 2017), and *M. chelonae* subsp. *gwanakae* (Kim et al., 2018), have both types of MSADs, MSAD-1 and MSAD-2, in their genomes (**Supplementary Table 1**). Of note, overt slow-grower human pathogens, including *M. tuberculosis, M. leprae*, *Mycobacterium marinum, M. ulcerans*, and *M. avium* subsp. *paratuberculosis,* do not have any MSAD genes in their genomes (**Supplementary Table 1** and **Figure 3**). In addition, of 49 slow-growing strains with an MSAD gene, all of those from the Tuberculosis-Simiae clade (emended genus *Mycobacterium*) have the MSAD-1 type but not the MSAD-2 type in their genomes. However, all members of the Terrae clade (*Mycolicibacter* gen. nov.) and Triviale clade (*Mycolicibacillus* gen. nov.) have MSAD-2 but not MSAD-1. Our finding of no MSAD genes in the genomes of more than half the *Mycobacterium* species, particularly in overt human pathogens, suggests gene loss during the evolutionary adaptation of *Mycobacterium* towards host-associated lifestyles, particularly in slow-growing human pathogens. In addition, the finding that two slow grower groups, the Tuberculosis-Simiae clade and the *M. terrae* complex group, including the Terrae and Triviale clades, have distinct MSAD types, MSAD-1 and MSAD-2, respectively (**Supplementary Table 1** and **Figure 3**), suggests the distinct roles of the two MSAD types in slow-grower evolution.

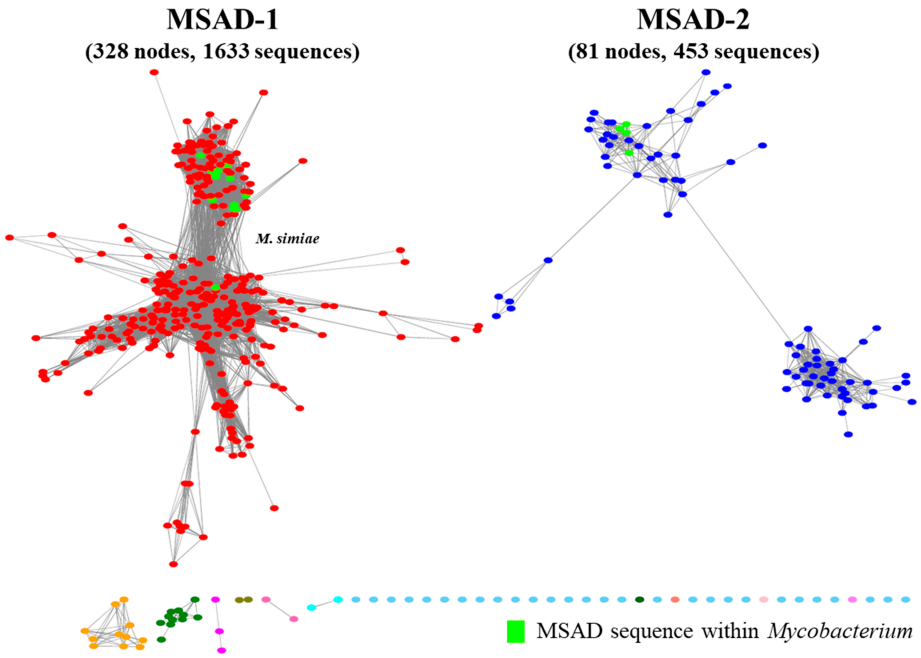


FIGURE 2
A new malonate semialdehyde decarboxylase (MSAD) protein similarity network including mycobacterial MSADs constructed in this study. Visualized protein similarity network of MSADs within the biosphere. Total MSAD family protein sequences (2,078 sequences) and mycobacterial MSADs (100 sequences) were used to construct the network and the network was visualized through the EFI-EST webserver. MSADs consist of two independent branches, MSAD-1 (red) and MSAD-2 (blue). The MSAD-1 branch consists of 328 nodes with 1,633 sequences. The MSAD-2 branch consists of 81 nodes with 453 sequences. All mycobacterial MSADs were scattered among the MSAD-1 and MSAD-2 groups.

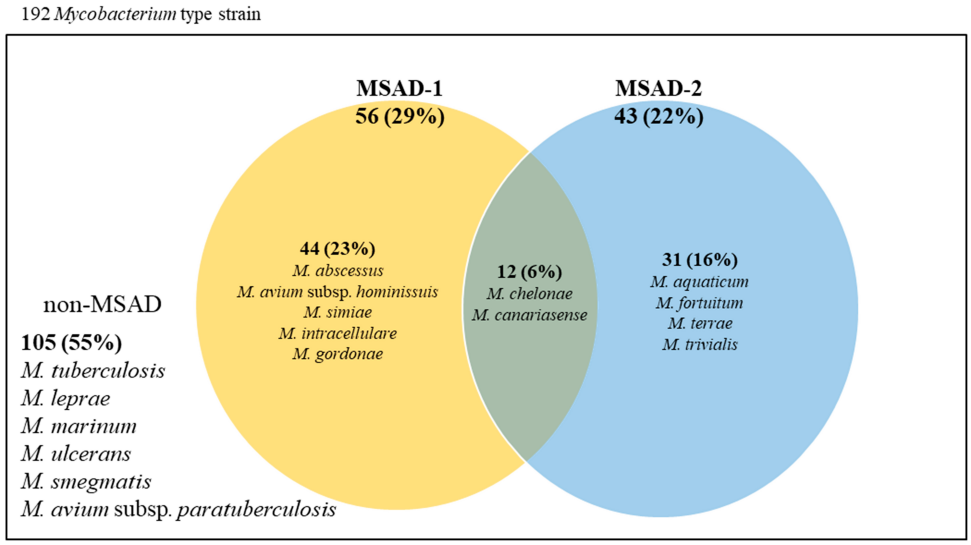


FIGURE 3
Venn diagram representing the distributions of MSAD-1 and MSAD-2 among strains within the genus *Mycobacterium*. Of all 192 *Mycobacterium* reference strains, only 87 (45.3%) strains have MSAD genes in their genome. Of these, 74 strains have either MSAD-1 or MSAD-2 genes in their genome, and 12 strains, including three from subspecies of *M. chelonae*, have both MSAD-1 and MSAD-2 genes. The remaining 105 strains, including strains of *M. tuberculosis* and *M. leprae*, do not have MSAD-like genes.

3.3. Distinct primary structures between MSAD-1s of slow-grower, MSAD-1s of rapid-grower and MSAD-2s

Our protein similarity network analysis indicated that all 100 mycobacterial MSADs consisted of a total of 17 nodes, with 13 nodes for MSAD-1s and 4 nodes for MSAD-2s (**Figure 2**). The predicted structures of both MSAD types of *M. chelonae* were identified as consistent with the MSAD model (PDB FG41) with 100% confidence, despite low protein percent identity between strains of MSAD-1 and MSAD-2 (75 and 25%, respectively). Both MSAD-1s and MSAD-2s share the characteristic structures of MSADs such as the β-α-β structure and proline-1 sequences of the TSF signature (**Figure 4A**; Davidson et al., 2018).

TABLE 1 Distribution of MSAD-1 among strains of *Mycobacterium avium* subspecies.

Subspecies name	MSAD protein sequence similarity (%)	Number of strains	
		MSAD	Non-MSAD
Mycobacterium avium	98.7–100	121	134
M. avium subsp. *avium*	99–100	17	26
M. avium subsp. *hominissuis*	98.7–100	104	58
M. avium subsp. *paratuberculosis*		0	49
M. avium subsp. *silvaticum*		0	1

None of the *M. avium* subspecies *paratuberculosis* and *silvaticum* strains have MSAD-1. Even among *M. avium* subspecies *avium* and *hominissuis*, the presence or absence of MSAD varies based on the strain of the species.

Tertiary structure prediction was performed for several *Mycobacterium* species according to each type. One of the template MSADs originating from *P. pavonaceae* (PpMSAD) have critical amino acids that facilitate binding to substrates (Poelarends et al., 2003, 2004, 2005; Almrud et al., 2005). Briefly, the Pro-1 and β-α-β structure shape is required for MSAD enzyme activity. Asp-37 and a pair of arginines, Arg-73 and Arg-75, are thought to act as linkers in the enzyme–substrate complex. Trp-114, Phe-116, Phe-123, and Leu-128 also participate in the stabilization of the substrate by forming a hydrophobic wall in the MSAD homotrimer (Almrud et al., 2005). Meanwhile, one of the other template MSAD originated from *C. bacterium* strain fg41 (FG41 MSAD) has a distinct primary structure compared with that of PpMSAD (Poelarends et al., 2008b; Guo et al., 2013). Despite also having Pro-1, a β-α-β structure shape, and Asp-37, similar to PpMSAD, FG41 MSAD has different signature protein sequences, indicating that the mechanisms of the two MSADs are different. Side chains of Thr-72, Gln-73, Arg-76, and Tyr-123 replace the pair of arginine residues in FG41 MSAD (**Figure 4B**).

Our data showed that all 57 mycobacterial MSAD-1 proteins have Pro-1, Asp-37, Trp-114, Phe-116, and Leu-128 (**Figure 5**). However, there are differences between rapid-growers and slow-growers in the substrate interaction region. While the MSAD-1 of rapid-growers has Thr-72 and Gln-73 except in *Mycobacterium aubagnense* and *Mycobacterium phocaicum*, the MSAD-1s of slow-growers have the Arg-73 and Arg-75 pair without exception, similar to PpMSAD (**Figures 5A, B**).

We also found that mycobacterial MSAD-2 has a distinct primary structure compared with that of MSAD-1. All MSAD-2 groups also start with Pro-1, and most of them have Asp-37 (**Figure 5C**). However, the MSAD-2 of *Mycobacterium koreense*, *Mycobacterium parakoreense*, and *Mycobacterium trivialis* belonging to the Triviale clade encodes Asn-37 instead of Asp-37. Moreover, unlike MSAD-1, MSAD-2 has no other signature

10.3389/fmicb.2023.1275616

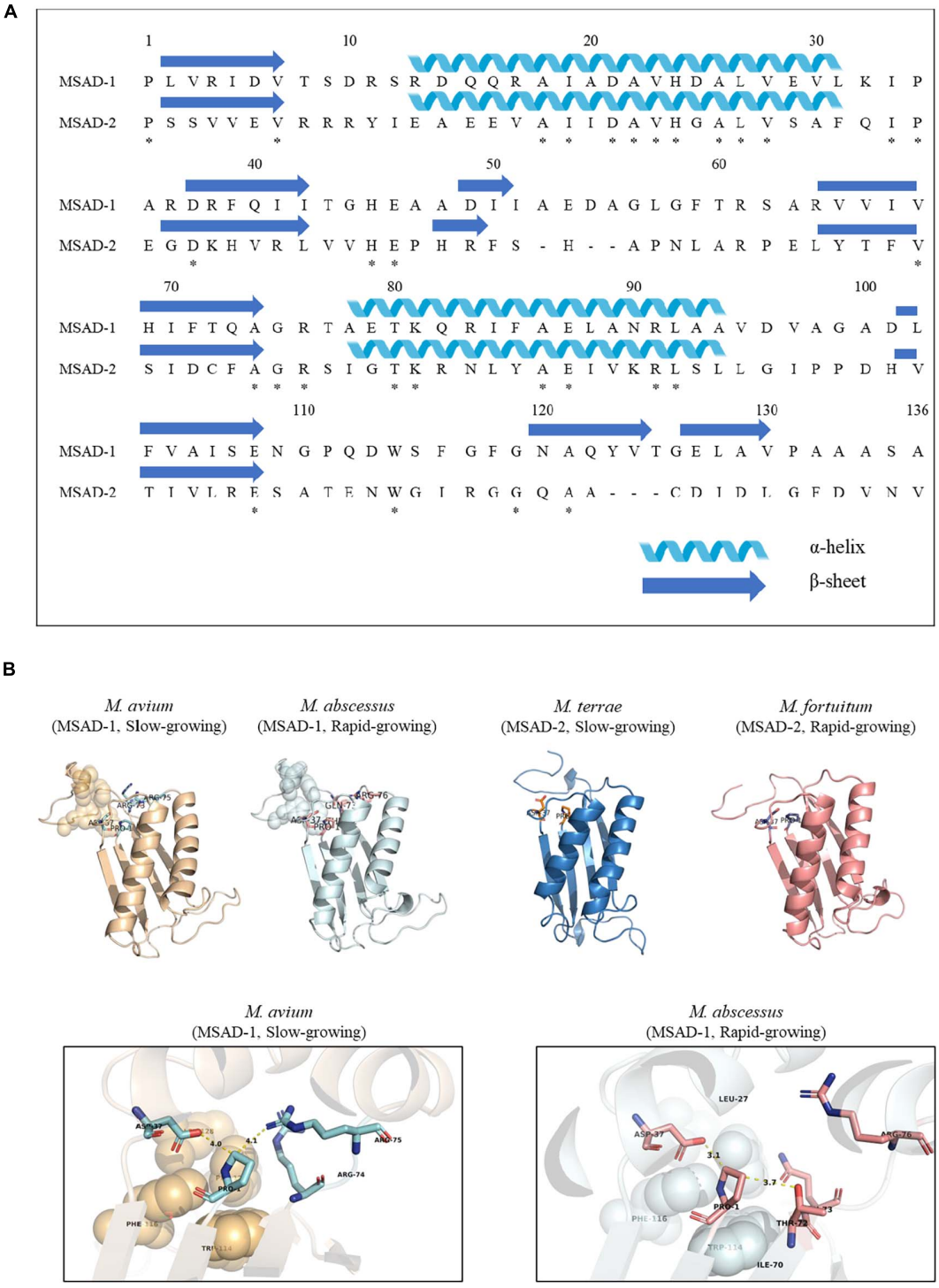


FIGURE 4
The primary and tertiary structures of mycobacterial MSAD-1 and MSAD-2 proteins. This number corresponds to the codon number of the *P. pavonaceae* MSAD protein. **(A)** Primary structure of MSAD-1 and MSAD-2 of *M. chelonae* ATCC 35752^{T} (MSAD-1: NCBI accession number OLT81431.1 and MSAD-2: OLT83119.1). The arrow indicates the β-sheet structure, while the spiral indicates the α-helix structure. **(B)** Ribbon diagram representing predicted tertiary structures of MSAD-1 of *M. avium* (slow-grower, left), MSAD-1 of *M. abscessus* (rapid-grower, center left), MSAD-2 of *M. terrae* (slow-grower, center right), and MSAD-2 of *M. fortuitum* (rapid-grower, right). MSAD-1 types showed a hydrophobic wall (shown as sphere) behind the active site (shown as stick) of Pro-1 and Asp-37. The structure of MSAD-2 is the same as MSAD-1 in that the positions of Pro-1 and Asp-37 located. *Identical amino acids between two MSADs.

sequences, including the pair of arginines (Arg-73 and Arg-75) or sequences related to the hydrophobic wall (Figure 5C). Our data showed that sequences of MSAD-2 are more conserved between strains (61.6–100%) than those of the MSAD-1 group (29.7–100%) (Supplementary Figure 1A). Together, our data indicated that there are three distinct primary protein structures of MSADs

10.3389/fmicb.2023.1275616

A

MSAD-1	Species	1	...	37	...	72	73	74	75	76	...	114	115	116	...	123	...	128
MSAD reference sequence	*P. pavonaceae*	P		D		S	R	P	R	S		W	S	F		F		L
Slow growing Mycobacterium	*M. simiae*	P	...	D	...	S	R	R	R	T	...	W	S	F	...	F	...	L
	M. paragordonae	P		D		S	R	R	R	P		W	S	F		F		L
	M. intracellulare chimaera	P		D		S	R	R	R	P		W	S	F		F		L
	M. intracellulare yongonense	P		D		S	R	R	R	P		W	S	F		F		L
	M. intracellulare intracellulare	P		D		S	R	R	R	P		W	S	F		F		-
	M. avium hominissuis	P		D		S	R	R	R	P		W	S	F		F		L
		*		*			*		*			*		*		*		**

B

MSAD-1	Species	1	...	37	...	72	73	74	75	76	...	114	115	116	...	123	...	128
MSAD reference sequence	*C. bacterium*	P		D		T	Q	A	G	R		W	S	F		Y		L
Rapid growing Mycobacterium	*M. chelonae bovis*	P	...	D	...	T	Q	T	G	R	...	W	S	F	...	Y	...	L
	M. chelonae gwanakae	P		D		T	Q	T	G	R		W	S	F		Y		L
	M. chelonae	P		D		T	Q	A	G	R		W	S	F		Y		L
	M. abscessus bolletii	P		D		T	Q	T	G	R		W	S	F		Y		L
	M. abscessus massiliense	P		D		T	Q	T	G	R		W	S	F		Y		L
	M. abscessus	P		D		T	Q	T	G	R		W	S	F		Y		L
		*		*		*	*			*		*		*		*		*

C

MSAD-2	Species	1	...	37	...	72	73	74	75	76	...	114	115	116	...	123	...	128
Slow growing Mycobacterium	*M. terrae*	P	...	D	...	G	R	S	L	D	...	R	G	G	...	E	...	V
	M. algericus	P		D		G	R	T	L	E		R	G	G		E		V
	M. trivalis	P		G		G	R	S	L	Q		R	G	G		D		V
Rapid growing Mycobacterium	*M. fortuitum*	P		D		G	R	S	V	Q		R	G	G		D		V
	M. chelonae	P		D		G	R	S	I	G		R	G	G		D		V
	M. chelonae gwanakae	P		D		G	R	S	I	G		R	G	G		D		V
	M. chelonae bovis	P		D		G	R	S	I	G		R	G	G		D		V
		*		>														

FIGURE 5
Distinct primary structures between mycobacterial MSAD-1 of slow-growers and rapid-growers and MSAD-2 of slow-growers and rapid-growers. This number corresponds to the codon number of the *P. pavonaceae* MSAD protein (PpMSAD). **(A)** The primary structure of MSAD-1 of slow-growers. MSAD-1 of slow-growers has structural characteristics similar to those of *P. pavonaceae* MSAD (PpMSAD), such as a pair of arginine (Arg-73 and Arg-75) residues and a hydrophobic wall (Trp-114, Phe-116, Phe-123, and Leu-128). **(B)** The primary structure of MSAD-1 of rapid-growers. MSAD-1 of rapid-growers has structural characteristics similar to those of *Coryneform bacterium* MSAD (FG41 MSAD), such as side chains (Thr-72, Gln-73, and Tyr-123) and a hydrophobic wall (Trp-114, Phe-116, and Leu-128). **(C)** The primary structure of MSAD-2. MSAD-2 does not have any specific sequence related to the enzymatic reaction of MSAD, regardless of the growth characteristics of *Mycobacterium*. *Indicates that all type strains have the same nucleotide. **Means that most of them have the same nucleotide. >Is the same nucleotide except for the trivalis clade.

between *Mycobacterium* strains (MSAD-1 of slow-growers, MSAD-1 of rapid-growers, and MSAD-2), suggesting their distinct roles in the evolution and pathogenesis of *Mycobacterium*.

3.4. Phylogenetic analysis based on MSAD-1 and MSAD-2 sequences

To assess the phylogenetic relationships of MSAD-1s and MSAD-2s among *Mycobacterium* strains, we performed phylogenetic analysis based on mycobacterial MSAD-1 and MSAD-2 sequences. First, the mycobacterial MSAD-1 phylogenetic tree was constructed from DNA sequences of 57 mycobacterial MSAD-1s from 56 *Mycobacterium* strains (two independent MSAD-1s in *M. simiae*) with DNA lengths ranging from 348 to 411 bp. The G+C content of MSAD-1 ranges from 56.1 to 68.4%. We found that most *Mycobacterium* strains can be separated at the species level, showing sequence similarity levels ranging from 29.7 to 100% (**Supplementary Figure 1A**). In general, the MSAD-1 DNA sequence-based phylogenetic tree revealed natural relationships between *Mycobacterium*, as shown in the *hsp65*-based tree (**Supplementary Figure 2**), clearly including separation between slow-growers (Tuberculosis-Simiae clade) and rapid-growers and separation between the Fortuitum-Vaccae clade and Abscessus-Chelonae clade. Of note, our MSAD-1-based phylogenetic analysis showed that the MSAD of five strains (*M. simiae, M. dioxanotrophicus, Mycobacterium farcinogenes, Mycobacterium senegalense,* and *Mycobacterium agri*) did not belong to the *Mycobacterium* clade, suggesting that their MSAD-1s was laterally transferred from another bacterial group (**Figure 6A**). In parallel, our further BLASTn analysis also supports LGT transfer of their MSAD-1 genes (**Supplementary Table 3**). Phylogenetic analysis based on MSAD-1 protein sequences also showed a topology similar to that based on MSAD-1 DNA sequences (**Supplementary Figure 3**).

MSAD-2 phylogenetic analysis was performed using 43 DNA sequences of mycobacterial MSAD-2, with DNA sequence lengths ranging from 357 to 435 bp. The G+C content of MSAD-2 is slightly higher than that of MSAD-1, ranging from 58 to 69.5%, and most strains can be separated at the species level, showing sequence similarity levels ranging from 61.6 to 100%, indicating that MSAD-2 is more conserved than MSAD-1 (**Supplementary Figure 1B**). Our MSAD-2 phylogenetic analysis also reveals natural relationships between mycobacterial strains, including separation between four clades, the Fortuitum-Vaccae and Abscessus-Chelonae clades of rapid-growers and

10.3389/fmicb.2023.1275616

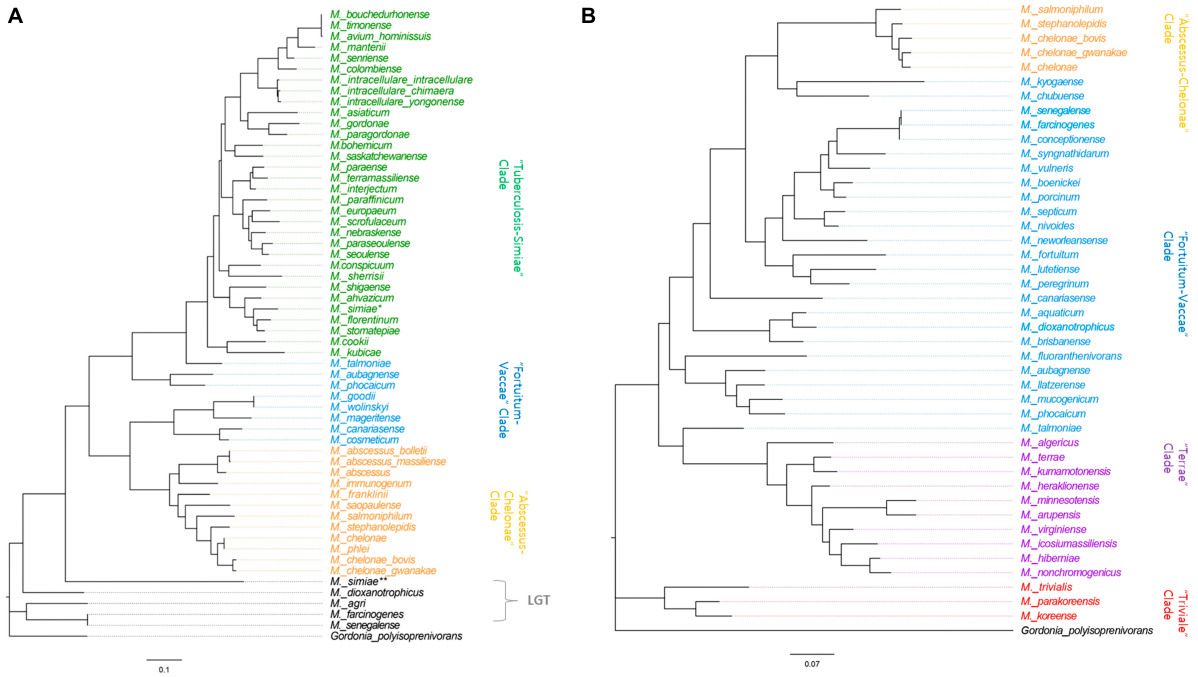


FIGURE 6
Phylogenetic trees based on the MSAD-1 (348–411 bp) and MSAD-2 (357–435 bp) sequences. **(A)** The mycobacterial MSAD-1 phylogenetic tree was constructed from DNA sequences of 57 mycobacterial MSAD-1s from 56 mycobacterial strains (two independent MSAD-1s in *M. simiae*) with DNA lengths ranging from 348 to 411 bp. **(B)** The MSAD-2 phylogenetic analysis was constructed using 43 DNA sequences of mycobacterial MSAD-2s, with DNA sequence lengths ranging from 357 to 435 bp.

Terrae and Triviale clades of slow-growers (Figure 6B), which is also shown in the MSAD-2 protein-based phylogeny (Supplementary Figure 3B). Together, our findings show that MSAD-1 and MSAD-2 sequences basically reflect the phylogenetic relationships between strains within the genus *Mycobacterium* except for some strains subject to LGT of their MSAD-1 gene, suggesting their pivotal role in the pathogenesis and evolution of *Mycobacterium* speciation.

3.5. MSAD-1 distribution between strains of *Mycobacterium avium* subspecies

Mycobacterium avium (Ma) is one of the most virulent NTM species, causing a broad spectrum of diseases in humans and ruminant animals as a member of the MAC, and it consists of four subspecies, namely, *M. avium* (Maa), *M. hominissuis* (Mah), *M. paratuberculosis* (Map), and *M. silvaticum* (Mas) (Thorel et al., 1990). Due to their distinct pathogenic potentials, their subspecies separation has recently gained great attention (Turenne et al., 2007). In this study, to investigate MSAD distributions between strains of Ma subspecies, a total of 255 genome sequences of four Ma subspecies were analyzed using tBLASTn. Of note, none of the 49 Map strains and 1 Mas strain had the MSAD gene in their genomes (Table 1). Only some Maa strains and Mah strains have MSAD-1 orthologs in their genomes. Of the 43 Maa strains, only 17 strains (39.53%) have MSAD-1 genes, showing 99.0–100% sequence similarity values between strains. In particular, a type strain of Maa, ATCC 25291^T, does not have MSAD in its genome. In the case of Mah, 104 of 162 strains (64.19%) have MSADs in their genomes, showing 98.7–100% sequence similarity values between strains (Table 1). Of note, among the 121 Ma strains with MSADs in their genomes, 13 (11 from Mah and 2 from Maa) have identical types of mutations in their MSAD genes, including a total of three mutations, two types of silent mutations, C81T and G87T, and a one-letter deletion at site 98T, which causes premature termination of the MSAD protein (Figure 7). Interestingly, 9 of 13 Ma strains with truncated MSADs were isolated from domestic pigs in Japan (Table 2; To et al., 2020; Komatsu et al., 2021), highlighting their potential roles in the pathogenesis or epidemiology of swine mycobacteriosis.

Together with our finding that there are no MSAD orthologs in the genome of Map strains, the finding that 13 of the 121 Ma strains with an MSAD-1 gene in their genome have mutated MSAD genes with premature translation termination further highlights the role of MSAD pseudogenization in species or subspecies differentiation of Ma.

3.6. Induction of pro-inflammatory cytokines in murine macrophage, J774A.1 cell by MSAD-1 protein of *Mycobacterium abscessus*

Our findings showing that loss of MSAD gene in the genome of pathogenic mycobacteria during their evolutionary adaptation prompt us to hypothesize that MSAD loss could contribute to their pathogenesis via evading immune responses of innate cells,

10.3389/fmicb.2023.1275616

ssp.	strain		79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	208	209	210	211	212	213	214
P. pavonaceae	2AAJ	Nucleotide	A	T	G	G	T	C	G	A	T	G	C	A	T	T	C	G	G	C	G	T	C	C	C	T	G	T	G	A	T	A	T
		Protein		M			V			D			A			F			G			V			P			V			I		
Mah	TH135 (Type strain)	Nucleotide	G	T	C	G	T	C	G	A	G	G	C	G	T	T	C	G	G	C	G	T	T	C	C	G	G	T	G	G	T	G	A
		Protein		V			V			E			A			F			G			V			P			V			V		
Mah	MAH-E-106	Nucleotide	G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	GM5		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	Toy195		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	Toy194		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU486		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU483		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU480		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU477		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU474		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	OCU471		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	LYM086		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	FLAC0146		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
Mah	MAV 061107 1842		G	T	T	G	T	C	G	A	T	G	C	G	T	T	C	G	G	C	G	-	T	C	C	G	G	T	G	G	T	G	A
		Protein		V			V			E			A			F			G			V			R			W			*		
			*	*	>	*	*	*	*	*	>	*	*	*	*	*	*	*	*	*	*	>>	*	*	*	*	*	*	*	*	*	*	*	*

FIGURE 7
MSAD-1 sequence alignments of 13 *Mycobacterium avium* strains with truncated MSADs. The 13 *M. avium* strains with truncated MSADs have mutations at C81T and G87T and a one-letter gene deletion at 98T. The MSAD protein of *M. avium* consists of 130 amino acids, but the 13 strains above were truncated to 71 amino acids due to frameshift by deletion. The symbol "*" indicates that all mycobacterium nucleotides are the same. The symbol ">" represents the site where synonymous mutation occurred and the symbol ">>" where deletion occurs and a frameshift occurs. The black arrow indicates a skip from the 102nd nucleotide to the 208th nucleotide. This number corresponds to the codon number of the *P. pavonaceae* MSAD protein.

macrophage, or dendritic cells. To address this hypothesis, we evaluated effect of MSAD-1 of *M. abscessus* on production of pro-inflammatory cytokines in murine macrophage, J774A.1 cells (**Figure 8**). We found that MSAD-1 treatment exerted enhanced productions of two inflammatory cytokines, TNF-α and IL-6 in J774A.1 cells in a dose dependent manner, suggesting that it could elicit immune response of innate cells in infection of mycobacteria.

4. Discussion

Comparative genomic analysis has revealed reductive genome evolution over the course of mycobacterial speciation, particularly in pathogenic slow-growing strains with strict host-associated lifestyles, including *M. tuberculosis* (Gagneux and Small, 2007; ten Bokum et al., 2008; Gagneux, 2018), *M. leprae* (Gómez-Valero et al., 2007; van Hooij and Geluk, 2021), and *M. ulcerans* (Stinear et al., 2007; Röltgen et al., 2012). Despite its extensive distribution throughout the biosphere, MSAD is not a housekeeping gene; thus, its distribution between mycobacterial strains with broad-spectrum lifestyles from environmental saprophytes to strict host-restricted pathogens could provide a deep understanding of their evolution and pathogenesis.

There are several noteworthy findings in this study. First, we found that there are no MSAD orthologs in the genome of more than half of the analyzed reference mycobacterial strains (105 of 192 strains, 54.7%) (**Figure 3**). In particular, highly virulent strains such as *M. tuberculosis*, *M. leprae*, *M. ulcerans, M. marinum*, Map, and *Mycobacterium kansasii,* which are regarded as overt human pathogens, do not have MSADs in their genomes (**Supplementary Table 1**), consistent with previous findings of their reductive genome evolution (Brosch et al., 2001; Veyrier et al., 2011; Wee et al., 2017). Moreover, we did not find any MSAD orthologs in any strains of Map, which is more virulent than other Ma subspecies, including Maa and Mah (Turenne et al., 2007) (**Figure 7**). In addition, we confirmed pseudogenization events in the MSAD genes of 13 strains of Maa and Mah by one-letter gene deletion causing frameshift events (**Figure 7**), further supporting our hypothesis of MSAD gene loss in more pathogenic mycobacteria.

Second, our SSN and phylogenetic analysis revealed two distinct MSAD types, MSAD-1 and MSAD-2, in *Mycobacterium* (**Figures 2**–**5**). We found that in 87 of 192 strains (45.3%), MSAD-1s (found in 57 sequences from 56 slow-growing strains and rapid-growing strains, two distinct MSAD-1s in *M. simiae*) and MSAD-2s (found in 43 strains in the *M. terrae* complex and rapid-growing strains) were distributed in an exclusive manner

TABLE 2 Summary information of 13 *M. avium* strains with truncated MSAD-1 genes.

Subspecies	Strain	Accession_NCBI	Genome length (bp)	Accession_assembly	Country	Patient	CDS	Reference
hominissuis	MAH-E-106	NZ_LMVZ00000000	5111701	GCF_001610545.1	Germany	Soil	4,876	Unpublished
hominissuis	GM5	NZ_WEGX00000000	5037010	GCF_020055295.1	Japan	Swine	4,844	Komatsu et al., 2021
hominissuis	Toy194	NZ_WEGP00000000	5347524	GCF_020055155.1	Japan	Swine	5,001	
hominissuis	Toy195	NZ_WEGO00000000	5346468	GCF_020055125.1	Japan	Swine	5,013	
hominissuis	OCU480	NZ_WEHE00000000	5088946	GCF_020055405.1	Japan	Swine	4,792	
hominissuis	OCU474	NZ_WEHK00000000	5087878	GCF_020054555.1	Japan	Swine	4,792	
hominissuis	OCU477	NZ_WEHH00000000	5087664	GCF_020055495.1	Japan	Swine	4,749	
hominissuis	OCU486	NZ_WEGY00000000	5023805	GCF_020055345.1	Japan	Swine	4,703	
hominissuis	OCU471	NZ_WEHN00000000	4990913	GCF_020055565.1	Japan	Swine	4,683	
hominissuis	OCU483	NZ_WEHB00000000	4943024	GCF_020055385.1	Japan	Swine	4,624	
hominissuis	LYM086	NZ_FKJM00000000	5304361	GCF_900079545.1	Unknown	Unknown	5,105	Unpublished
avium	FLAC0146	NZ_NSFH00000000	5179834	GCF_002292595.1	United States	Homo sapiens	4,875	Unpublished
avium	MAV_061107_1842	NZ_JAOQ00000000	5320946	GCF_000523655.1	United States	Homo sapiens	4,981	Unpublished

Of these, nine strains were isolated from domestic pigs in Japan.

among *Mycobacterium* species. However, 12 rapid-growing strains, including three *M. chelonae* members, had both types of MSADs, MSAD-1 and MSAD-2, in their genomes (**Supplementary Table 2**). These findings suggest an evolutionary event from a common ancestor with both types of MSADs, possibly belonging to the rapid-growing group, into the strain with either a single MSAD-1 or a single MSAD-2 during *Mycobacterium* speciation. Both types of MSADs show distinct distributions among *Mycobacterium* strains. In particular, among slow-growing strains with an MSAD gene, the Tuberculosis-Simiae clade, including most pathogenic slow-growing strains, has only MSAD-1 (**Figure 6**), suggesting a pivotal role of MSAD-1 in the pathogenesis and evolution of pathogenic slow-grower. In contrast, the Triviale and Terrae clades have only the MSAD-2 type in their genomes (**Figure 6**), highlighting the role of MSAD-2 in their pathogenesis and evolution. In rapid-growers, the MSAD distribution is more complex than that in slow-grower. Rapid-growing strains with the MSAD gene could be divided into three groups according to their MSAD distributions: a group with only the MSAD-1 type, including three subspecies of *M. abscessus*, namely, *M. abscessus*, *M. massiliense* and *M. bolleti*, a group with only the MSAD-2 type, including most of the Fortuitum-Vaccae clade, and a group with both types of MSADs, MSAD-1 and MSAD-2, including three members of the *M. chelonae* subspecies (**Figure 6** and **Supplementary Table 1**), suggesting distinct roles of the two MSAD types in pathogenesis and evolution in terms of respective rapid-growing groups. Given the close phylogenetic relationships between *M. chelonae* and *M. abscessus* strains (Adékambi and Drancourt, 2004; Tortoli et al., 2017), it is tempting to speculate on MSAD-2 loss during speciation from the common ancestor of two species with both types of MSADs into *M. abscessus*, which may have contributed to its pathogenesis.

Third, our data showed that the MSAD-1s of the slow-growers and rapid-growers and MSAD-2s have distinct primary signature sequences that can play a crucial role in their function. The MSAD-1s of all slow-growers have Arg-73, Arg-75, and Asp-37, which play a critical role in substrate reactions, and Trp-114, Phe-116, Phe-123, and Leu-128, which are essential for the hydrophobic wall, as shown in *P. pavonaceae* strain 170's MSAD (PpMSAD) (**Figure 5A** and **Supplementary Data 1**; Almrud et al., 2005), suggesting a similar role of the MSADs of slow-growers with that of the latter. On the other hand, the MSAD-1s of rapid-growers have a primary signature sequence of Thr-72, Gln-73, and Tyr-123 (Guo et al., 2013) instead of a pair of arginines reacting with the substrate, distinct from the pattern in slow-growers, as shown in the MSADs of *C. bacterium* strain FG41 (FG41 MSAD) (**Figure 5B** and **Supplementary Data 1**; Guo et al., 2013), suggesting distinct functional roles or distinct evolutionary selection pressures between the MSAD-1s of slow-growers and rapid-growers. Interestingly, all the strains with MSAD-2 have primary sequences distinct from MSAD-1, suggesting distinct functional roles of MSAD-2s from slow-growers and rapid-growers MSAD-1s (**Figure 5C** and **Supplementary Data 1**). Further research is needed to determine the exact role of MSAD-2s in enzyme function.

Our phylogenetic analysis showed that MSAD-2 is more resistant than MSAD-1 to LGT events (**Figure 6**). Comparing the results with phylogenetic trees based on MSAD-1, MSAD-2, and *hsp65* sequences, we could not find any LGT events in 43 strains with MSAD-2, but 5 of 53 strains had MSAD-1s (9.4%)

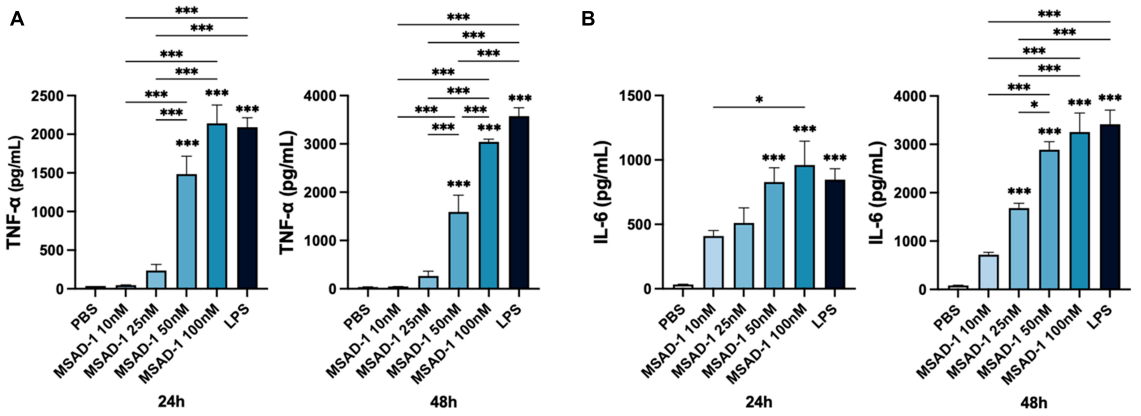


FIGURE 8
Induction of pro-inflammatory cytokines, TNF-α, and IL-6, in murine macrophage, J774A.1 cells, by MSAD-1 protein of *Mycobacterium abscessus*. **(A)** The levels of TNF-α induced by MSAD-1 were measured by ELISA in the culture medium after 24 or 48 h. **(B)** The levels of IL-6 induced by MSAD-1 were measured by ELISA in the culture medium after 24 or 48 h. Data represent means ± SEM (standard error of mean) from quadruplicate samples and are representative of at least two independent experiments. Statistical analysis was performed using one-way ANOVA with Tukey's multiple comparisons test, and statistical significance are denoted by asterisks (*$P < 0.05$, ***$P < 0.001$). An asterisk above the bar indicates statistical significance compared to the PBS control group, and an asterisk with a line indicates statistical significance compared to the indicated group.

(the second MSAD-1 type of *M. simiae, M. dioxanotrophicus, M. farcinogenes, M. senegalense,* and *M. agri*), which may have been laterally transferred from another bacterial group (Figure 6A and Supplementary Table 3). Of these, 4 strains (*M. dioxanotrophicus, M. farcinogenes, M. senegalense,* and *M. agri*) belong to the rapid-growers of the Fortuitum-Vaccae clade. Interestingly, we found that while one (accession number BBX43605.1) of two MSAD-1 genes of *M. simiae* belongs to the clade of Tuberculosis-Simiae by MSAD-1-based phylogenetic analysis, the other MSAD-1 of *M. simiae* (accession number BBX40959.1) did not (Figure 6A). Our protein similarity network analysis indicated that it has sequence similarity with the MSAD from the *Methylobacterium phylum* (Supplementary Table 3), suggesting acquisition by *M. simiae* of the second MSAD-1 gene via LGT.

Our *in vitro* experiment using MSAD-1 protein indicated that MSAD-1 could evoke inflammatory response from innate cells in mycobacteria infection, suggesting that MSAD-1 loss in pathogenic mycobacteria could contribute into their chronic infection or pathogenesis via evading immune response of innate cell. However, the role of two distinct MSADs, MSAD-1 and MSAD-2, in mycobacterial pathogenesis or evolution must be proved in the future via further *in vitro* and *in vivo* studies using MSAD gene knock out or reinforced mutant.

5. Conclusion

In conclusion, our data revealed two distinct types of MSADs, MSAD-1 and MSAD-2, among strains in the *Mycobacterium* genus, but more than half of the strains, including strains of pathogenic mycobacteria such as *M. tuberculosis*, *M. leprae, M. marinum, M. ulcerans* and Map, have no MSAD orthologs in their genomes. Furthermore, in 13 *Ma* strains, MSAD-1 pseudogenization was found, suggesting MSAD-1 loss during host adaptation of pathogenic mycobacteria. Loss of MSAD during speciation could contribute to their pathogenicity via escape from host innate immune cells.

There are several limitations to this study. Study regarding the role of MSAD in mycobacteria evolution and pathogenesis is mainly focused on the bioinformatics prediction. Biochemical and structural evidence based on actual enzyme activities of MSAD-1 and MSAD-2 have not been introduced. So, these limitations should be addressed in the future.

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

B-JK: Conceptualization, Data curation, Funding acquisition, Project administration, Validation, Writing – original draft, Writing – review and editing. DL: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. DK: Data curation, Methodology, Software, Validation, Visualization, Writing – review and editing. HS: Conceptualization, Formal analysis, Writing – review and editing. SC: Data curation, Methodology, Validation, Visualization, Writing – review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study was supported by grant from the Seoul National University Hospital (SNUH) Research Fund

(Grant No. 03-2020-0180) and Korea Health Industry Development Institute (KHIDI), funded by Ministry of Health and Family Welfare, Republic of Korea (Grant No. HI22C0312). The funder was not involved in the study design, data analysis and writing of the manuscript.

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.2023.1275616/full#supplementary-material

References

Abe, Y., Fukushima, K., Matsumoto, Y., Niitsu, T., Nabeshima, H., Nagahama, Y., et al. (2022). *Mycobacterium senriense* sp. nov., a slowly growing, non-scotochromogenic species, isolated from sputum of an elderly man. *Int. J. Syst. Evol. Microbiol.* 72:005378. doi: 10.1099/ijsem.0.005378

Adékambi, T., and Drancourt, M. (2004). Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing. *Int. J. Syst. Evol. Microbiol.* 54, 2095–2105. doi: 10.1099/ijs.0.63094-0

Akiva, E., Brown, S., Almonacid, D., Barber, A. II, Custer, A., Hicks, M., et al. (2014). The structure–function linkage database. *Nucleic Acids Res.* 42, D521–D530. doi: 10.1093/nar/gkt1130

Almrud, J., Poelarends, G., Johnson, W., Serrano, H., Hackert, M., and Whitman, C. (2005). Crystal structures of the wild-type, P1A mutant, and inactivated malonate semialdehyde decarboxylase: A structural basis for the decarboxylase and hydratase activities. *Biochemistry* 44, 14818–14827. doi: 10.1021/bi051383m

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J, Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bouam, A., Armstrong, N., Levasseur, A., and Drancourt, M. (2018). *Mycobacterium terramassiliense*, *Mycobacterium rhizamassiliense* and *Mycobacterium numidiamassiliense* sp. nov., three new *Mycobacterium simiae* complex species cultured from plant roots. *Sci. Rep.* 8:9309. doi: 10.1038/s41598-018-27629-1

Brosch, R., Pym, A., Gordon, S., and Cole, S. (2001). The evolution of mycobacterial pathogenicity: Clues from comparative genomics. *Trends Microbiol.* 9, 452–458. doi: 10.1016/S0966-842X(01)02131-X

Chalmers, J., Aksamit, T., Carvalho, A., Rendón, A., and Franco, I. (2018). Non-tuberculous mycobacterial pulmonary infections. *Pulmonology* 24, 120–131. doi: 10.1016/j.pulmoe.2017.12.005

Dahl, J., Gatlin, I. W., Tran, P., and Sheik, C. (2021). *Mycolicibacterium nivoides* sp. Nov. isolated from a peat bog. *Int. j. Syst. Evol. Microbiol.* 71:004438. doi: 10.1099/ijsem.0.004438

Davidson, R., Baas, B., Akiva, E., Holliday, G., Polacco, B., LeVieux, J., et al. (2018). A global view of structure–function relationships in the *Tautomerase* superfamily. *J. Biol. Chem.* 293, 2342–2357. doi: 10.1074/jbc.M117.815340

Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* 16, 202–213. doi: 10.1038/nrmicro.2018.8

Gagneux, S., and Small, P. (2007). Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect. Dis.* 7, 328–337. doi: 10.1016/S1473-3099(07)70108-1

Gómez-Valero, L., Rocha, E., Latorre, A., and Silva, F. (2007). Reconstructing the ancestor of *Mycobacterium leprae*: The dynamics of gene loss and genome reduction. *Genome Res.* 17, 1178–1185. doi: 10.1101/gr.6360207

Guo, Y., Serrano, H., Poelarends, G., Johnson, W. Jr., Hackert, M., and Whitman, C. (2013). Kinetic, mutational, and structural analysis of malonate semialdehyde decarboxylase from Coryneform bacterium strain FG41: mechanistic implications for the decarboxylase and hydratase activities. *Biochemistry* 52, 4830–4841. doi: 10.1021/bi400567a

Gupta, R., Lo, B., and Son, J. (2018). Phylogenomics and comparative genomic studies robustly support division of the genus Mycobacterium into an emended genus Mycobacterium and four novel genera. *Front. Microbiol.* 9:67. doi: 10.3389/fmicb.2018.00067

Gutierrez, M., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., et al. (2005). Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. *PLoS Pathog* 1:e5. doi: 10.1371/journal.ppat.0010005

He, Y., Mathieu, J., Yang, Y., Yu, P., da Silva, M., and Alvarez, P. (2017). 1, 4-Dioxane biodegradation by Mycobacterium dioxanotrophicus PH-06 is associated with a group-6 soluble di-iron monooxygenase. *Environ. Sci. Technol. Lett.* 4, 494–499. doi: 10.1021/acs.estlett.7b00456

Huddleston, J., Burks, E., and Whitman, C. (2014). Identification and characterization of new family members in the tautomerase superfamily: analysis and implications. *Arch. Biochem. Biophys.* 564, 189–196. doi: 10.1016/j.abb.2014.08.019

Jankute, M., Cox, J., Harrison, J., and Besra, G. (2015). Assembly of the mycobacterial cell wall. *Annu. Rev. Microbiol.* 69, 405–423. doi: 10.1146/annurev-micro-091014-104121

Jeong, H., Lee, S., Seo, H., Kim, D., Lee, D., and Kim, B. (2022). Potential of Mycobacterium tuberculosis chorismate mutase (Rv1885c) as a novel TLR4-mediated adjuvant for dendritic cell-based cancer immunotherapy. *Oncoimmunology* 11:2023340. doi: 10.1080/2162402X.2021.2023340

Johansen, M., Herrmann, J., and Kremer, L. (2020). Non-tuberculous mycobacteria and the rise of Mycobacterium abscessus. *Nat. Rev. Microbiol.* 18, 392–407. doi: 10.1038/s41579-020-0331-1

Jones, P., Binns, D., Chang, H., 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

Kelley, L., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kim, B., Kim, B., Jeong, J., Lim, J., Park, S., Lee, S., et al. (2018). A description of Mycobacterium chelonae subsp. gwanakae subsp. Nov., a rapidly growing mycobacterium with a smooth colony phenotype due to glycopeptidolipids. *Int. J. Syst. Evol. Microbiol.* 68, 3772–3780. doi: 10.1099/ijsem.0.003056

Kim, B., Kim, G., Kim, B., Jeon, C., Jeong, J., Lee, S., et al. (2017). Description of Mycobacterium chelonae subsp. bovis subsp. nov., isolated from cattle (*Bos taurus* coreanae), emended description of Mycobacterium chelonae and creation of Mycobacterium chelonae subsp. chelonae subsp. nov. *Int. J. Syst. Evol. Microbiol.* 67, 3882–3887. doi: 10.1099/ijsem.0.002217

Kim, H., Kim, S., Shim, T., Kim, M., Bai, G., and Park, Y. (2005). Differentiation of Mycobacterium species by analysis of the heat-shock protein 65 gene (hsp65). *Int. J. Syst. Evol. Microbiol.* 55, 1649–1656. doi: 10.1099/ijs.0.63553-0

Komatsu, T., Ohya, K., Ota, A., Nishiuchi, Y., Yano, H., Matsuo, K., et al. (2021). Genomic features of Mycobacterium avium subsp. hominissuis isolated from pigs in Japan. *Gigabyte* 2021:gigabyte33. doi: 10.46471/gigabyte.33

Lancaster, E., Yang, W., Johnson, W. Jr., Baas, B., Zhang, Y., and Whitman, C. (2022). Kinetic, Inhibition, and Structural Characterization of a Malonate Semialdehyde Decarboxylase-like Protein from Calothrix sp. PCC 6303: A Gateway

to the non-Pro1 Tautomerase Superfamily Members. *Biochemistry* 61, 1136–1148. doi: 10.1021/acs.biochem.2c00101

Lotti, T., and Hautmann, G. (1993). Atypical mycobacterial infections: a difficult and emerging group of infectious dermatoses. *Int. J. Dermatol.* 32, 499–501. doi: 10.1111/j.1365-4362.1993.tb02832.x

Marri, P., Bannantine, J., and Golding, G. (2006). Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. *FEMS Microbiol. Rev.* 30, 906–925. doi: 10.1111/j.1574-6976.2006.00041.x

Nishiuchi, Y., Iwamoto, T., and Maruyama, F. (2017). Infection sources of a common non-tuberculous mycobacterial pathogen. Mycobacterium avium complex. *Front. Med.* 4:27. doi: 10.3389/fmed.2017.00027

Oberg, N., Zallot, R., and Gerlt, J. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J. Mol. Biol.* 2023:168018. doi: 10.1016/j.jmb.2023.168018

Petrini, B. (2006). Mycobacterium abscessus: an emerging rapid-growing potential pathogen. *Apmis* 114, 319–328. doi: 10.1111/j.1600-0463.2006.apm_390.x

Poelarends, G., Johnson, W., Murzin, A., and Whitman, C. (2003). Mechanistic characterization of a bacterial malonate semialdehyde decarboxylase: identification of a new activity in the tautomerase superfamily. *J. Biol. Chem.* 278, 48674–48683. doi: 10.1074/jbc.M306706200

Poelarends, G., Serrano, H., Johnson, W., Hoffman, D., and Whitman, C. (2004). The hydratase activity of malonate semialdehyde decarboxylase: mechanistic and evolutionary implications. *J. Am. Chem. Soc.* 126, 15658–15659. doi: 10.1021/ja044304n

Poelarends, G., Serrano, H., Johnson, W., and Whitman, C. (2005). Inactivation of malonate semialdehyde decarboxylase by 3-halopropiolates: Evidence for hydratase activity. *Biochemistry* 44, 9375–9381. doi: 10.1021/bi050296r

Poelarends, G., Serrano, H., Person, M., Johnson, W. Jr., and Whitman, C. (2008a). Characterization of Cg10062 from Corynebacterium glutamicum: implications for the evolution of cis-3-chloroacrylic acid dehalogenase activity in the tautomerase superfamily. *Biochemistry* 47, 8139–8147. doi: 10.1021/bi8007388

Poelarends, G., Veetil, V., and Whitman, C. (2008b). The chemical versatility of the β–α–β fold: Catalytic promiscuity and divergent evolution in the tautomerase superfamily. *Cell. Mol. Life Sci.* 65, 3606–3618. doi: 10.1007/s00018-008-8285-x

Poelarends, G., and Whitman, C. (2004). Evolution of enzymatic activity in the tautomerase superfamily: mechanistic and structural studies of the 1, 3-dichloropropene catabolic enzymes. *Bioorgan. Chem.* 32, 376–392. doi: 10.1016/j.bioorg.2004.05.006

Qi, W., Käser, M., Röltgen, K., Yeboah-Manu, D., and Pluschke, G. (2009). Genomic diversity and evolution of Mycobacterium ulcerans revealed by next-generation sequencing. *PLoS Pathog* 5, e1000580. doi: 10.1371/journal.ppat.1000580

Ratnatunga, C., Lutzky, V., Kupz, A., Doolan, D., Reid, D., Field, M., et al. (2020). The rise of non-tuberculosis mycobacterial lung disease. *Front. Immunol.* 11:303. doi: 10.3389/fimmu.2020.00303

Röltgen, K., Stinear, T., and Pluschke, G. (2012). The genome, evolution and diversity of Mycobacterium ulcerans. *Infect. Genet. Evol.* 12, 522–529. doi: 10.1016/j.meegid.2012.01.018

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., 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

Stinear, T., Seemann, T., Pidot, S., Frigui, W., Reysset, G., Garnier, T., et al. (2007). Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *Genome Res.* 17, 192–200. doi: 10.1101/gr.5942807

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

ten Bokum, A., Movahedzadeh, F., Frita, R., Bancroft, G., and Stoker, N. (2008). The case for hypervirulence through gene deletion in Mycobacterium tuberculosis. *Trends in microbiology.* 16, 436–441. doi: 10.1016/j.tim.2008.06.003

Thorel, M., Krichevsky, M., and Vincent Lévy-Frébault, V. (1990). Numerical taxonomy of mycobactin-dependent mycobacteria, emended description of Mycobacterium avium, and description of Mycobacterium avium subsp. avium subsp. nov., Mycobacterium avium subsp. paratuberculosis subsp. nov., and Mycobacterium avium subsp. silvaticum subsp. nov. *Int. J. Syst. Evol. Microbiol.* 40, 254–260. doi: 10.1099/00207713-40-3-254

To, K., Cao, R., Yegiazaryan, A., Owens, J., and Venketaraman, V. (2020). General overview of nontuberculous mycobacteria opportunistic pathogens: *Mycobacterium avium* and *Mycobacterium abscessus*. *J. Clin. Med.* 9, 2541. doi: 10.3390/jcm9082541

Tortoli, E., Fedrizzi, T., Meehan, C., 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

Turenne, C., Wallace, R. Jr., and Behr, M. (2007). Mycobacterium avium in the postgenomic era. *Clin. Microbiol. Rev.* 20, 205–229. doi: 10.1128/CMR.00036-06

van Hooij, A., and Geluk, A. (2021). In search of biomarkers for leprosy by unraveling the host immune response to Mycobacterium leprae. *Immunol. Rev.* 301, 175–192. doi: 10.1111/imr.12966

Veyrier, F., Dufort, A., and Behr, M. (2011). The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol.* 19, 156–161. doi: 10.1016/j.tim.2010.12.008

Vincent Lévy-Frébault, V., and Portaels, F. (1992). Proposed minimal standards for the genus Mycobacterium and for description of new slowly growing Mycobacterium species. *Int. J. Syst. Evol. Microbiol.* 42, 315–323. doi: 10.1099/00207713-42-2-315

Wee, W., Dutta, A., and Choo, S. (2017). Comparative genome analyses of mycobacteria give better insights into their evolution. *PLoS One.* 12, e0172831. doi: 10.1371/journal.pone.0172831

Zallot, R., Oberg, N., and Gerlt, J. (2019). The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* 58, 4169–4182. doi: 10.1021/acs.biochem.9b00735

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 01 November 2023
DOI 10.3389/fmicb.2023.1285027

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Sonia Dávila-Ramos,
Autonomous University of the State of Morelos, Mexico
Chang-Wei Lei,
Sichuan University, China

*CORRESPONDENCE
Yan Su
✉ 2006au@163.com

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

RECEIVED 29 August 2023
ACCEPTED 10 October 2023
PUBLISHED 01 November 2023

CITATION
Zhang Y, Lv F, Su Y, Zhang H and Zhang B (2023) Complete genome sequencing and comparative genomic analysis of three donkey *Streptococcus equi* subsp. *equi* isolates.
Front. Microbiol. 14:1285027.
doi: 10.3389/fmicb.2023.1285027

COPYRIGHT
© 2023 Zhang, Lv, Su, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Complete genome sequencing and comparative genomic analysis of three donkey *Streptococcus equi* subsp. *equi* isolates

Yuwei Zhang†, FenFen Lv†, Yan Su*, Huan Zhang and Baojiang Zhang

Department of Microbiology and Immunology, College of Veterinary Medicine, Xinjiang Agricultural University, Ürümqi, Xinjiang, China

Introduction: *Streptococcus equi* subspecies *equi* (*S. equi*) is the causative agent of strangles, which is one of the most common and highly contagious respiratory infectious illnesses in horses. *Streptococcus equi* (*S. equi*) is a horse-specific pathogen that originated from the closely related zoonotic pathogen *Streptococcus equi* subspecies *zooepidemicus* (*S. zooepidemicus*). Despite decades of research, the movement of genetic material across host-restricted diseases remains a mystery.

Methods: Three *S. equi* donkey isolates (HTP133, HTP232, and HT1112) were recently isolated from a strangles epidemic on donkey farms in China's Xinjiang Province. In this study, we performed a comprehensive comparative analysis of these isolates using whole genome sequencing and compared them to the published genomic sequences of equine strain *S. equi* 4047 to uncover evidence of genetic events that shaped the evolution of these donkey *S. equi* isolates' genomes.

Results: Whole genome sequencing indicated that both strains were closely related, with comparable gene compositions and a high rate of shared core genomes (1788-2004). Our comparative genomic study indicated that the genome structure is substantially conserved across three donkey strains; however, there are several rearrangements and inversions when compared to the horse isolate *S. equi* 4047. The virulence factors conveyed by genomic islands and prophages, in particular, played a key role in shaping the pathogenic capacity and genetic diversity of these *S. equi* strains. Furthermore, we discovered that the HT133 isolate had a strong colonization ability and increased motility; the HT1112 isolates had a significantly higher ability for antimicrobial resistance and biofilm formation, and the HT232 isolate gained pathogenic specialization by acquiring a bacteriophage encoding hyaluronate lyase.

Discussion: In summary, our findings show that genetic exchange across *S. equi* strains influences the development of the donkey *S. equi* genome, offering important genetic insights for future epidemiological studies of *S. equi* infection.

KEYWORDS

***Streptococcus equi*, whole genome sequencing, comparative genomics, virulence, antimicrobial resistance, donkey**

Introduction

Streptococcus equi subspecies *equi* (*S. equi*) is the causative agent of strangles, which is one of the most commonly diagnosed and feared serious, highly contagious respiratory infectious equids diseases in the world, with morbidity rates of up to 100% (Waller, 2013; Cordoni et al., 2015). Outbreaks impose a significant financial burden on the equine industry and horse owners, as well as raising concerns about horse health and welfare (Boyle et al., 2018).

As a result, initiatives to enhance diagnosis and prevention are critical to the equestrian sector. The use of molecular biological approaches has tremendously benefited epidemiological research, such as tracking the epidemic (Bertelli and Greub, 2013). Furthermore, next-generation sequencing methods like whole genome sequencing (WGS) and comparative genomic analysis have recently become very strong tools for examining the increase and loss of particular virulence and antibiotic resistance genes. Gene gain by horizontal acquisition is a critical element in the establishment of novel pathogenic *streptococci* strains (Harris et al., 2015). Receiving bacteria may gain an instant selective advantage by acquiring antibiotic resistance mechanisms or producing novel virulence activities. Furthermore, such research provides information on the proclivity of certain isolates for invasion.

Previous studies have contributed to our understanding of the genetic and pathogenic characteristics of a couple of *S. equi* strains isolated from horses. Holden et al. (2009) sequenced the complete genome sequences of *S. equi* 4047 and *S. zooepidemicus* H70, analyzed their general genome features, and compared the genomes from around the world to uncover genetic evidence of their evolution events. Researchers (Alber et al., 2005; Paillot et al., 2010) found that the evolution of *S. equi* from an ancestral *S. zooepidemicus* is associated with the acquisition of four prophage-encoded virulence genes: *seeI*, *seeL*, *seeM*, and *seeH*. So far, only a limited number of genomic sequences of equine *S. equi* strains have been sequenced, and genomic resources for donkey isolates are still not available. It is not yet known how the donkey isolate affects the evolution and exchange of genetic material of these host-restricted pathogens.

Whole genome sequencing (WGS) and a subsequent comparative genomic analysis can be used to understand the movement of virulence and other host adaptation genes between *streptococci* isolates from different hosts. Furthermore, WGS results could help the prevention and control of strangles outbreaks. Thus far, there is no publicly available data regarding the sequences of *S. equi* donkey isolates recovered in China, which has limited our comprehensive and systematic understanding of these isolates' evolution and fitness in donkey hosts.

According to the 16S rRNA gene and physiological and biochemical studies, we have found three *S. equi* strains designated *S. equi* HTP133, HTP232, and HT1112 (ST-179) recovered from donkeys with strangles in Xinjiang, China, in 2020. We report on the WGS, virulence factors, and antibiotic susceptibility of three *S. equi* donkey isolates from Xinjiang, China, in order to investigate genetic diversity and virulence features. Following that, we compared the genomes of these donkey strains isolated in China to the genomes of *S. equi* 4047, a virulent strain isolated from a horse with strangles in the New Forest, England, in 1990 (Kelly et al., 2006), to better understand the evolution of within-host adaptation and shed new light on the evolution of this donkey *S. equi* lineage toward host restriction.

Materials and methods

S. equi isolates and culture conditions

Isolates of *S. equi* Lancefield group C were represented by Se HT1112 (2019), Se HTP133 (2020), and Se HTP232 (2020). *S. equi* was isolated from donkey samples of different farms during strangles outbreak in the Chinese region of Xinjiang and has been typed as ST-179 by multi-locus sequence typing (MLST). At 37°C, all strains were cultivated overnight in Todd-Hewitt Broth (THB) containing 5% sheep blood (OXOID, Basingstoke, Hampshire, England).

Genomic DNA extraction

The genomic DNA from *S. equi* isolates was extracted from bacterial cell pellets using a TIANamp Bacteria DNA Kit (Tiangen, Beijing, China) according to the manufacturer's instructions. DNA was resuspended in nuclease-free water and quantified with a NanoVue spectrophotometer (GE Healthcare, Little Chalfont, United Kingdom). DNA quality and RNA contamination were assessed by electrophoresis with 0.8% agarose gels.

Whole genome sequencing, functional genome annotation, and comparative genomic analysis

The whole genome was sequenced at Shanghai Majorbio Bio-pharm Technology Co., Ltd., (Shanghai, China) using a

single-molecule real-time (SMRT) sequencing platform, PacBio RS II (version 2.3.0, Pacific Biosciences, USA), and Illumina sequencing platforms.

The clean data were assembled using SOAP denovo2 for draft genomes and Canu or SPAdes v.3.8.0 for complete genomes. The quality-controlled reads were *de novo* assembled using the RS Hierarchical Genome Assembly Process protocol by SMRT Analysis. Genome annotation coding sequence (CDS) predictions and annotations were performed using RAST (Overbeek et al., 2014), Glimmer v3.02 (Delcher et al., 1999), and the predicted gene sequences were translated and aligned against the National Center for Biotechnology Information Non-Redundant Protein Database (NR) database, the Gene Ontology database, the Clusters of Orthologous Groups (COG) (von Mering et al., 2003), Swiss-Prot, Pfam, and the Kyoto Encyclopedia of Genes and Genomes database (Kanehisa et al., 2012; Kanehisa and Goto, 2000). The circular genome map was constructed with CGView Server (Grant and Stothard, 2008). Genomic islands (GIs) were predicted by IslandViewer 4 (Bertelli et al., 2017).

PHASTER (Arndt et al., 2016) and IslandViewer 4 (Bertelli et al., 2017) were used to identify prophages and gene islands, respectively. The CRISPR Recognition Tool v1.1 (Bland et al., 2007) was used to predict CRISPR. The Circos program was used to generate circular genomes based on anticipated ORFs, rRNA, tRNA, prophages, gene islands, and GC skew information (Krzywinski et al., 2009). Stothard and Wishart (2005) used CGView to create the genome atlas. Resistance Gene Identifier software was used to annotate ARDs (Antibiotic Resistance Datas) against the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2020).

The reference sequences of these three *S. equi* strains (HT1112, HTP133, and HTP232) were the entire genome sequences of strain *S. equi* 4047, and the BLAST Ring Image Generator (BRIG) (Alikhan et al., 2011) was used for genome-wide comparison to build a circular genomic map. Darling et al. (2010) used progressive Mauve to align the four *S. equi* strains' genomes. The Majorbio Cloud Platform 3′s online tool OrthoMCL 2 was utilized for comparative analysis to detect the common and unique genes contained by various isolates.

Pan-genome and core genome analysis

The pan-genome and core genome analyses were conducted as described by Medini et al. (2005). For pan-genome analyses, starting with a single genome, genomes were added in a randomized order without replacement at each fixed number of genomes. The sizes and numbers of the pan-genome and core genomes were calculated as a function of the number of genomes sequentially included.

Antimicrobial susceptibility testing

Antimicrobial susceptibility testing was performed by the disk diffusion method according to Clinical and Laboratory Standards Institute [CLSI] (2020) guidelines with Se HT1112, HTP133, and Se HTP232. Twenty-one antibiotics, including amoxicilin, ampicillin, cefuroxime, ceftiofur, cefoxitin, penicillin, gentamicin, streptomycin, erythromycin, clarithromycin, doxycycline, oxytetracycline, levofloxacin, tetracycline, norfloxacin, enrofloxacin, ciprofloxacin, sulfafurazole, sulfadiazine sodium, rifampin, and clindamycin, were selected to determine antimicrobial susceptibility. The *Escherichia coli* strain ATCC 25922 was used as the control strain for susceptibility studies.

Biofilm formation assay

The crystal violet technique was used for the biofilm experiment, as previously described (Wang et al., 2011; Zarankiewicz et al., 2012). An *S. equi* (colony HT1112, HTP133, and HTP232 strains) was injected into 5 mL of THB and cultured overnight at 37°C with shaking. Cultures were diluted 1:100 with new THB broth containing 1% fibrinogen and incubated at 37°C without shaking for 24 h. The positive control was *S. equi* spp. (Se HT1112, HTP133, and Se HTP232 strains) growing on a THB medium containing 1% fibrinogen. The negative control was uninoculated culture media containing 1% fibrinogen. The supernatant with unadhered bacteria was aspirated, and the wells were washed three times with 200 μL of sterile PBS. The wells were stained for 10 min with 200 μL 0.1% (w/v) crystal violet, then rinsed five times with 200 μL PBS to eliminate unbound crystal violet dye. Following the drying of the plate, biofilm-adsorbed crystal violet was resolubilized by adding 150 μL 95% (v/v) ethanol to each well and incubating for 15 min. The absorbance of each well was measured at 595 nm to evaluate biofilm levels. All tests were carried out in triplicate.

Virulence experiments and growth kinetics

Animal experiments were approved and conducted in strict compliance with the guidelines and regulations of the University of Xinjiang Agricultural University Animal Care and Use Committee.

Before inoculation of mice, cultures of Se HT1112, HTP133, and Se HTP232 strains were collected at 37°C until the logarithmic growth phase was reached, washed twice in THB, and adjusted to the appropriate dose. Experimental mice were injected subcutaneously with 5×10^8, 1×10^9, 2.5×10^9 CFU per mouse. Control mice were injected with sterile PBS. The mice were observed for 14 days for survival status, and the survival differences were plotted by Kaplan–Meier curves and evaluated using the log-rank test. The liver and spleen of experiment mice were homogenized and diluted properly for bacterial load investigation, and *S. equi* colonies were counted on TH medium.

A single colony of *S. equi* (Se HT1112, HTP133, and Se HTP232 strains) was injected into 5 mL of THB and cultured for 24 h with shaking at 37°C. To assess growth, the OD 600 was measured and recorded every 2 h. The sequences utilized for comparative genomic analysis were from *S. equi* subsp. *zooepidemicus* strain 4047 (FM204883) (Holden et al., 2009).

Evaluation of hyaluronidase activity

Cultures of Se HT1112, HTP133, and Se HTP232 were cultivated in the THB medium to mid-logarithmic phase.

Five milliliters of OD600 = 0.4 culture were centrifuged, the supernatants discarded, and the bacteria pellets resuspended in 100 μl of deionized water. The cells were then treated with 0.5% hyaluronic acid (HA) for 30 min at 40°C before being boiled for 3 min to inactivate them. Hyaluronidase activity was determined by adding 400 μL of DNS solution (Sigma), boiling for 5 min and measuring the optical density at 540 nm using a BioTek ELx808 Absorbance Microplate Reader (BioTek, Winooski, United States).

Nucleotide sequence accession number

The complete genome sequences of three donkey *S. equi* strain (HT1112, HTP133 and HTP232) have been submitted to the GenBank database with accession number CP133957, CP133955, CP133956.

Statistical analysis

Each experiment in this study has been performed for at least three biological replicates. Data were analyzed with the software SPSS version 16.0 and GraphPad Prism version 7 (GraphPad, La Jolla, CA, United States). Data are presented as mean ± SE or as geometric mean. For each assay, we applied a one-way ANOVA with Tukey's multiple comparisons test to assess the differences between groups. The survival data between mouse groups were analyzed with the log-rank test (Mantel–Cox). A value of $p < 0.05$ was considered statistically significant.

Results

Genome assembly and features of *S. equi* HT1112, HTP133, and HTP232

Features of the three newly sequenced and four published genomes of *S. equi* are shown in Table 1. The sizes of three *S. equi* genomes range from 2.15 Mb to 2.19 Mb, and the GC contents range from 41.30 to 41.36%. The predicted numbers of open reading frames (ORFs) range from 2,010 (1,824,747 bp, HTP133) to 2,223 (1,872,897 bp, HTP232), with average lengths ranging from 842 bp to 907 bp. The coding genes function (COG) analysis showed that the gene numbers of cellular component (823), molecular function (1,372) and biological process (1,160) in Se HT1112 were more than those of Se HTP133 (7,351,262,803) and HTP232 (7,441,264,806). The genomes of three isolates were examined for genome architecture, chromosomal sequences, and mobile elements such as GIs, prophage regions, and insertion elements (Figure 1). CRISPR is a unique family of direct repeat DNA sequences that widely exist in prokaryotic genomes. There are seven predicted CRISPR arrays in the Se HT1112 and HTP133 genomes and eight CRISPR arrays in the Se HTP232 strain. Features of the three newly sequenced genomes of *S. equi* are shown in Table 1.

Identification of prophages is important for the study of the genomes of the *S. equi* strains and their genetic potential. Prophage element analysis revealed that the genome of HTP232 harbors four prophages with sizes ranging from 8.6 to 17.5 kb, while the other two strains, HTP133 harbors three prophages and HT1112 harbors one prophage (Table 2). Seven *S. equi* and *S. zooepidemicus* isolates with different COG function classifications were selected for heatmap evaluation. Non-metric multidimensional scaling (NMDS) (Figure 2B) and heatmap (Figures 2A, C) based on COG function suggesting a high degree of relationship was observed for isolates HTP133, HT1112, HTP232, and these three donkey isolates have a close relationship with other two *S. equi* isolates 4047 and ATCC39506 (Figure 2C).

TABLE 1 Basic genomic characteristics of the strains.

Features	HT1112	HTP133	HTP232
Chromosome (bp)	2,156,526	2,150,351	2,191,031
Coding genes	2,167	2,010	2,223
Gene size (bp)	1,837,221	1,824,747	1,872,897
Gene average length (bp)	847.82	907.83	842.51
G + C content (%)	41.97	41.97	41.92
tRNA	77	67	67
rRNA	21	18	18
16s rRNA	7	6	6
23s rRNA	7	6	6
5s rRNA	7	6	6

Detection of putative virulence genes and comparison of growth

We found 169 (HT1112), 169 (HTP133), and 170 (HTP232) potential virulence factors (VFs) that were involved in 13 functions, including adherence, anti-phagocytosis, exoenzyme, invasion, iron uptake, regulation, complement protease, magnesium uptake, toxin, antiphagocytosis, serum resistance, stress protein, and secretion system. All three bacteria possess eight non-specific VFs that encode exoenzyme genes. When the VFs of the three strains were compared, we discovered that the exoenzyme-coding gene was only detected in the HTP232 strain and was not present in the other two isolates, HTP133 and HT1112 (Figure 3A).

In growth experiments, the growth rate dynamics revealed different growth rates among the three isolates. The *S. equi* HT1112 isolate had a shorter lag phase and a higher growth rate compared with the other two strains, HTP133 and HTP232 (Figure 3B).

Virulence-associated phenotypes

The connection between VFs and virulence-associated phenotypes of isolates HT1112, HTP133, and HTP232 was investigated to assess the virulence features. In a mouse infection paradigm, mice given the strains HT1112, HTP133, and HTP232 (1×10^9 CFU) all died within 24 h; the isolate HTP133 was the most deadly of the three. More germs were found in the spleen and lungs of mice infected with HTP133 than with HT1112 or HTP232,

10.3389/fmicb.2023.1285027

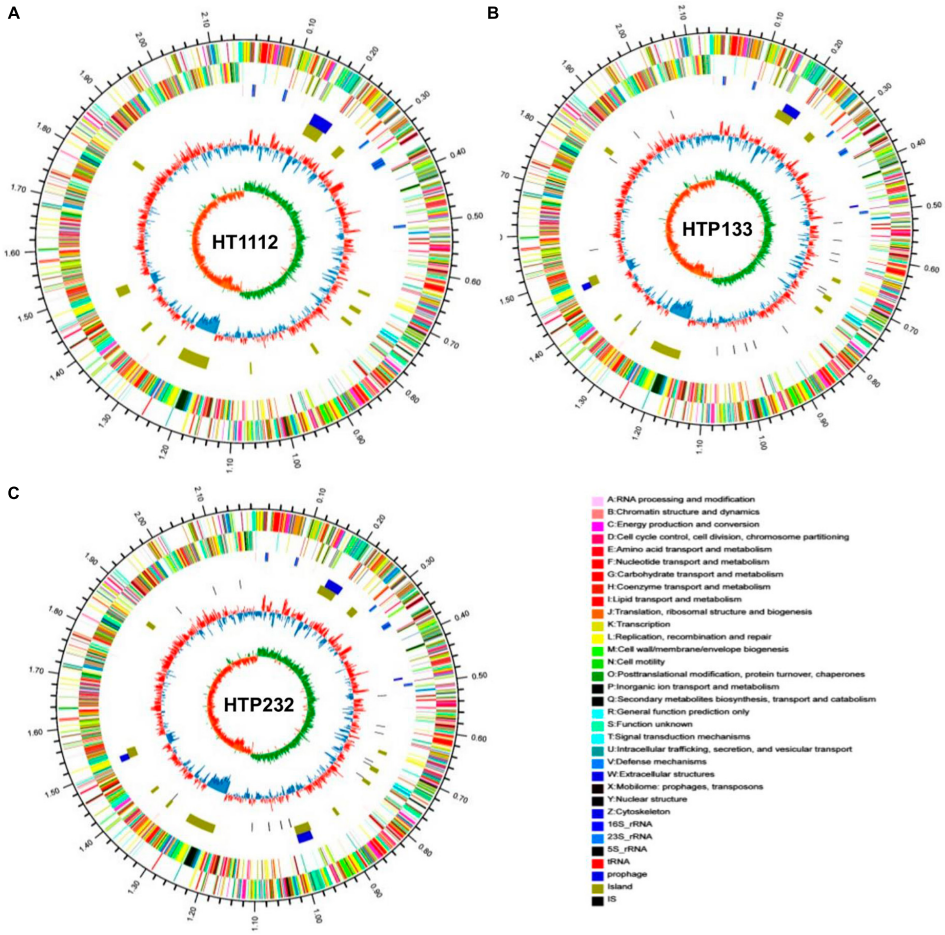


FIGURE 1
Schematic circular diagrams of three donkey *S. equi* isolates **(A)** HT1112, **(B)** HTP133, and **(C)** HTP232. The outer circles showed the scale marks of the genome. Circles 1 and 2 displayed the protein-coding genes on the forward strand and reverse strand, respectively. Circle 3 displayed the tRNA and rRNA genes on the forward strand. Circle 4 displayed the tRNA and rRNA genes. Circles 5, 6, and 7 displayed the prophage, genome island, and insert sequence, respectively. 16S rRNA (dark blue), 23S rRNA (pale blue), 5S rRNA (green), tRNA (red), prophage (purple), and island (yellow). The prophecies are displayed in blue. The brown color represents the genomic island. The genome map was made using Circos v0.64 (http://circos.ca/).

TABLE 2 **Annotated table of gene function results.**

Strains	GO	COG	KEGG	NR	Swiss-port	Prophage	Crispr-Cas	GI
HT1112	1,687	1,706	1,197	2,166	1,506	1	7	12
HTP133	1,527	1,708	1,202	2,010	1,509	3	7	10
HTP232	1,540	1,731	1,200	2,151	1,516	4	8	12

suggesting that this strain is particularly virulent and causes the most death and damage in animals (Figures 3C, D). Mice with HT1112, HTP133, or HTP232 infections had their lung and spleen tissues fixed for histopathological investigation (Figure 3E). The HE test showed that the lung tissues had been damaged by the challenge, showing enlarged alveolar septa, and the spleen had shown inflammation and bleeding.

Resistance genes and phenotypes

We used the CARD to investigate the genomes of the three isolates for antibiotic resistance genes (ARGs), and 172 resistance genes (of 22 types) were found there. These genes were potentially involved in resistance to macrolide, tetracycline, fluoroquinolone, peptide, lincosamide, pleuromutilin, streptogramin, aminoglycoside, glycopeptide, phenicol, oxazolidinone, acridine dye, nitroimidazole, rifamycin, fosfomycin, penam, diaminopyrimidine, aminoglycoside, mupirocin, sulfonamide, and glycylcycline. Among them, 25 tetracycline resistance genes and 44 macrolide resistance genes were discovered in the genomes of the samples (Figure 4A).

All of the isolates were subjected to antimicrobial susceptibility tests, and the results showed that a significant difference in resistance to antimicrobials was observed for these three strains. The HT1112 is resistant to six antimicrobials, including

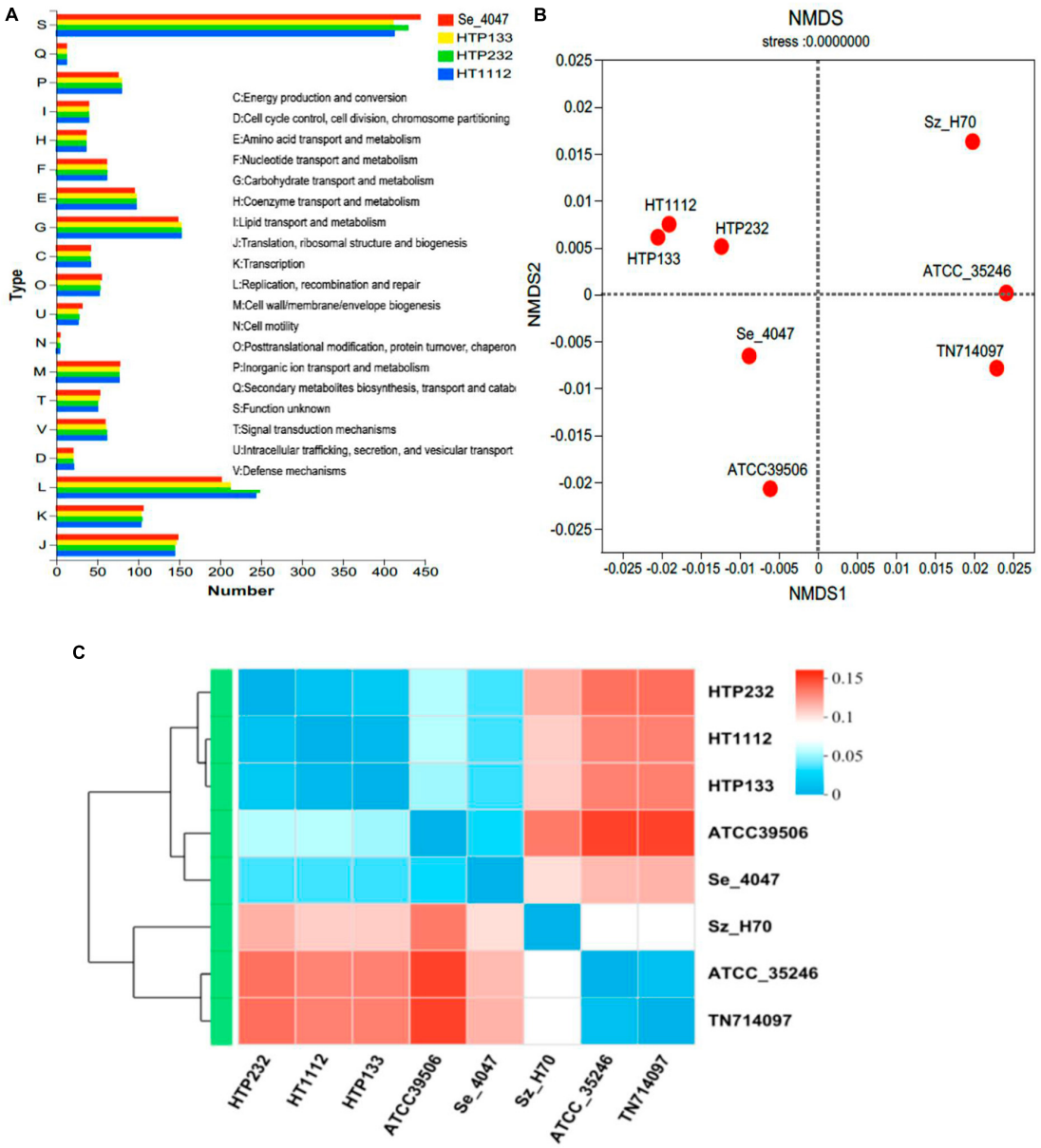


FIGURE 2
Analysis of the COG function of *S. equi* strains HT1112, HTP133, and HTP232 and comparison with the other 5 reference strains. **(A)** COG function comparison of HT1112, HTP133, and HTP232 with Se 4047 (GCF000026). **(B)** Non-metric multidimensional scaling (NMDS) analysis of three isolates (HT1112, HTP133, and HTP232 genomes) and five reference strain genomes based on COG function. The NMDS results revealed genome clustering in three isolates relative to these groups that were used as reference strains. **(C)** Heatmap based on the COG function showing evolution relations with respect to the COG function of six *streptococci* reference strain genomes and three isolates (HT1112, HTP133, and HTP232 genomes).

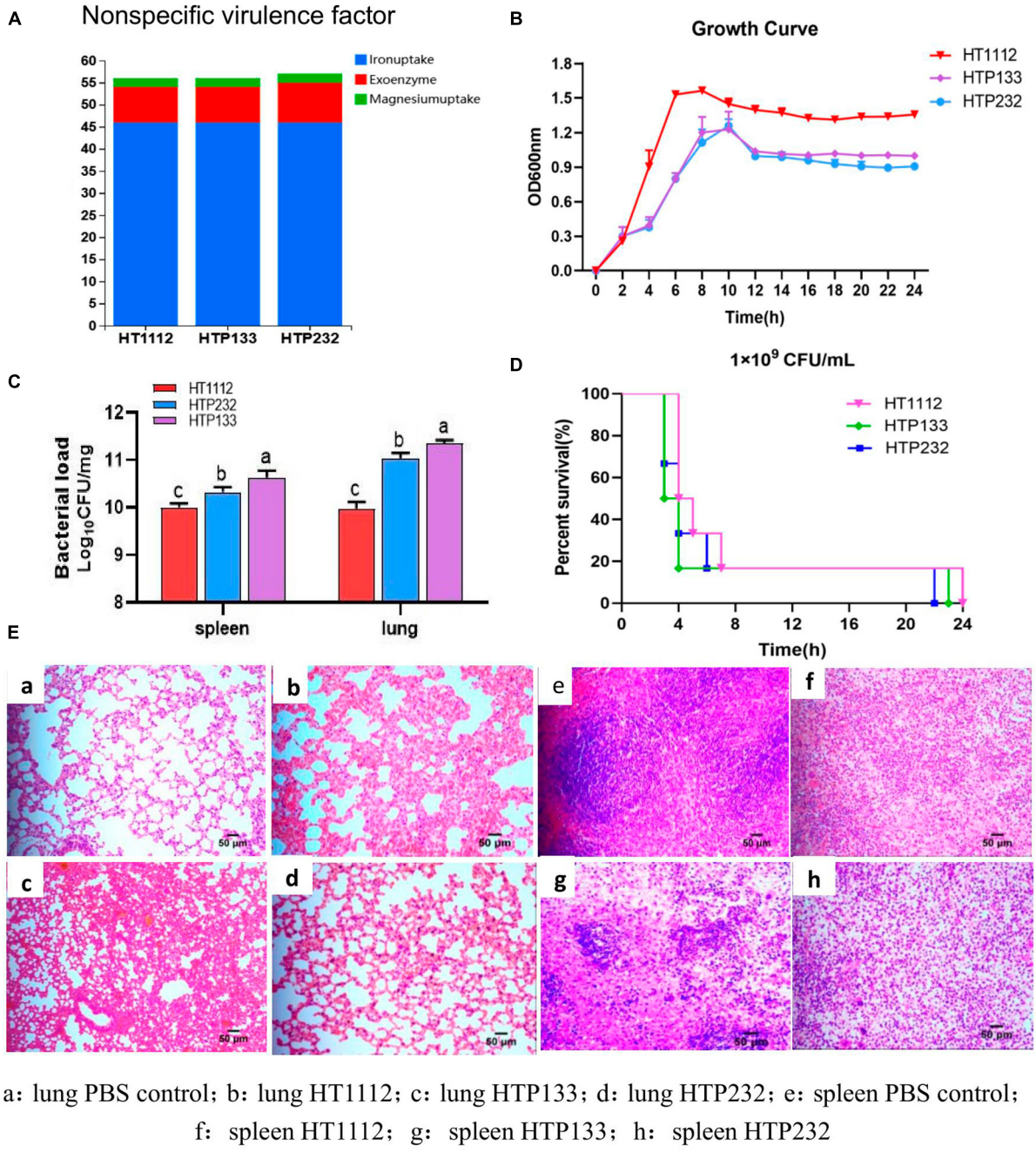
penicillin, cefoxitin, clarithromycin, ciprofloxacin, sulfadiazine sodium, and clindamycin, compared with two antimicrobials (cefuroxime and penicillin) for HTP133 and HTP232 isolates (Table 3).

Biofilm formation assay

Microtiter crystal violet tests were used to measure the amount of biofilm that these three *S. equi* isolates formed. Even though these three isolates formed biofilms differently from one another, Figure 4 shows a link between *S. equi* strains' biofilm production and antimicrobial potency. The HT1112 produces a lot of biofilms and has antibacterial properties (Figure 4B).

Comparative genomic analysis of HT1112, HTP133, and HTP232

We first observed a high degree of synteny was additionally for isolates HTP133, HT1112, and HTP232, suggesting a close relationship between these isolates (Figure 5). Genome synteny was further explored by comparing the linear organization of the chromosome of each *S. equi* isolate to *S. equi* 4047, and an increased



a: lung PBS control; b: lung HT1112; c: lung HTP133; d: lung HTP232; e: spleen PBS control; f: spleen HT1112; g: spleen HTP133; h: spleen HTP232

FIGURE 3
The virulence analysis of three isolates (HT1112, HTP133, and HTP232). **(A)** Distribution of non-specific virulence genes in three *S. equi* isolate genomes. **(B)** Growth curves of these three *S. equi* isolates. HT1112, HTP133, and HTP232 were cultured in THB broth at 37°C. The OD600 nm of the culture was measured every 2 h. The data represent the means and standard deviations of the results of three independent experiments. **(C)** Survival rates of *S. equi* strains, HT1112, HTP133, and HTP232, in BALB/C male mouse models. The groups of 5–6-week-old mice were subcutaneously inoculated with 1.0×10^9 CFU of *S. equi* isolates HT1112, HT232, and HT133 and monitored for 24 h. **(D)** At 24 h post-infection, the mice were sacrificed, and their spleens and lungs were harvested and homogenized to measure the CFUs of bacterial colonization. Colonies were expressed as Log10CFU/mg, and the results are shown as the mean of three mice in each group (SD, $p < 0.001$; $p < 0.01$; $p < 0.05$. CFU = colony-forming unit). The differences between survival curves were evaluated using the log-rank test ($n = 6$ mice). **(E)** Lung and spleen tissues of mice subjected to HE staining. There is a significant difference ($p < 0.05$) between numbers with different letters; no significant difference ($p > 0.05$) exists between numbers with the same letter.

number of genome inversions and rearrangements compared to *S. equi* 4047 were observed for isolates HT1112, HTP133, and HTP232 (Figures 5A, B, C). Since these isolates were from donkey cases, a high degree of global synteny and collinearity was observed among the HT1112 and HTP133 isolates (Figures 5D, E, F). Then we analyzed the shared and unique genes between isolates of HT1112, HTP133, HTP232, and *S. equi* 4047 (Figure 6). Core genomes are usually used to evaluate the genomic diversity within species. Our results showed 1,788 to 1,820 CDSs in the core genome were shared by these isolates; there were 69, 234, and 258 unique

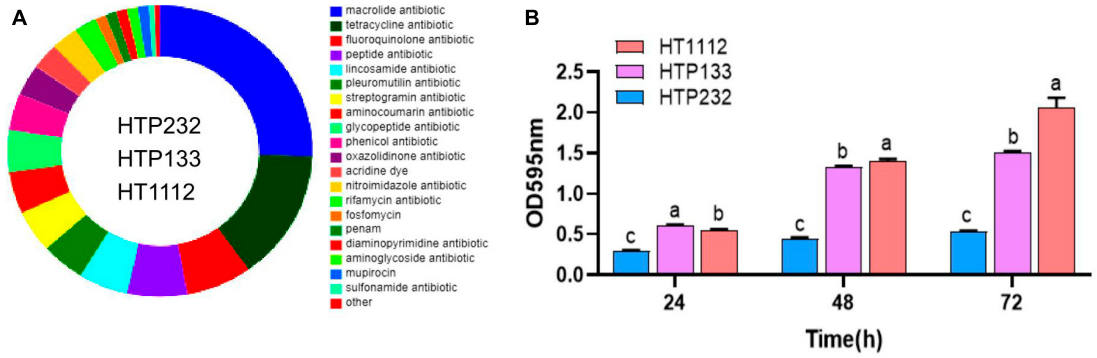


FIGURE 4
Analysis of the antibiotic resistance phenotype, resistance genes, and biofilm formation of three isolates (HT1112, HTP133, and HTP232). **(A)** Number and distribution of antibiotic resistance genes. **(B)** Biofilm formation by three isolates (HT1112, HTP133, and HTP232) determined by the microliter plate assay after crystal violet staining of bacterial cultures. The strain HT1112 showed higher biofilm formation than HTP133 and HTP232 ($P < 0.01$). The data represent the means and SDs of three independent experiments. There is a significant difference ($p < 0.05$) between numbers with different letters; no significant difference ($p > 0.05$) exists between numbers with the same letter.

genes harbored by the HTP133, HT1112, and HTP232 strains, respectively. On the other hand, we investigated the shared and unique genes between isolates of HT1112, HTP133, and HTP232 (**Figure 6A**). There were 1,828 to 2,004 shared genes among them, and there were 18 to 251 unique genes harbored by these isolates, respectively. Compared to *S. equi* 4047, donkey isolates, HTP133, HT1112, and HTP232, were found to contain a higher number of genome rearrangements and inversions.

Genomic islands (GIs) are fragments of DNA derived from horizontal gene transfer between different bacterial genomes (Rodriguez-Valera et al., 2016) and play an important role in the virulence of *S. equi*. In this study, we identified 13 GIs in the HTP232 genome (**Figure 6B**). Using the IslandPath-DIMOB program (Bertelli et al., 2017), and for the entire chromosome of strains HTP133 and HT1112, the GIs were 10 and 12, respectively.

Visualization of the alignment of the *S. equi* and the predicted CDS regions of the three strains, HT1112, HTP133, and HT P232, using the BRIG is given in **Figure 6C**. An increased number of genome rearrangements and inversions were observed for isolates and *S. equi* 4047.

Genome synteny was further explored by comparing the linear organization of the chromosomes of three *S. equi* isolates to *S. equi* 4047 using Mauve [14]. Genome structure, which is shown in **Figure 6D**, A high degree of global synteny was observed among the donkey isolates HT1112, HTP133, and HTP232, with the exception of one insertion in the HTP232 genome.

The alignment of each *S. equi* predicted CDS to their closest relatives, *S. equi* 4047, is shown in **Figure 6D**. Multiple genome alignments of the *S. equi* strains HT1112, HTP133, and HTP232 and *S. equi* 4047 identified one distinct block of the genome that only exists in HTP232 and *S. equi* 4047 strains, comprising the hyaluronidase gene cassette.

Hya on genomic island of *S. equi* isolates

The comparative genomic analysis revealed that the chromosomes of Se HTP133, Se HTP232, Se HT1112, and Se 4047 are generally collinear except for a genetic element transfer that we found between Se HTP232 and Se 4047 (**Figure 6D**). In Se 4047 GI13, the hyaluronate lyase-coding gene *hya* has been translocated to Se HTP232. Hyaluronate lyases are secreted enzymes that degrade HA and chondroitin, facilitating invasion by bacteria and their toxins. The transferred GI and hyaluronidase genes of Se 4047 were not conserved and consistent with those of the hyaluronidase genes in HTP232 (hyaluronidase gene, 1,112 and 1,883 bp) (**Figure 6E**).

Two CDS that code for hyaluronate lyase are present in the Se 4047-transferred GI 13 (26,322 bp). In addition, the two copies of hyaluronate lyase (761 and 1,883 bp) vary slightly in Se 4047. However, two distinct hyaluronidase genes (1,112 and 1,883 bp) are present in the transferred GI 13 (31,108 kb) of Se HTP232. Two distinct hyaluronate lyases have been acquired by Se HTP232 and are encoded on its GI 13 (**Figure 6E**).

To demonstrate that the hyaluronate lyase phenotype was due to a gene, we grew Se HTP133, Se HTP232, and the Se HT11127 isolate to test the hyaluronidase activity. Thus, the phenotypic profiling results indicated that the Se HTP232 isolate plays an important role in the hyaluronidase activity of Se HTP232 (**Figure 7**).

Discussion

Outbreaks of strangles in donkeys with severe clinical signs are becoming prevalent due to the increasing number of intensive donkey breeding farms in China (Dong et al., 2019). To date, there have been no published reports of genomic studies conducted on *S. equi* strains recovered from donkey farms in China. In this study, we determined the complete genome sequence of three donkey *S. equi* strains, HT1112, HTP133, and HTP232, isolated in China and used comparative genomic analysis to provide genetic evidence for their evolution. The whole genome sequence information of donkey-associated *S. equi* strains will contribute to the understanding of genetic features such as antimicrobial resistance and virulence-related genes.

TABLE 3 Antimicrobial resistance phenotype and antimicrobial resistance genes (ARG) of three strains.

Drugs	Phenotype (HT1112)	ARG	Phenotype (HTP133)	ARG	Phenotype (HT232)	ARG
Amoxicilin	S	PBP1a/PBP2x/ soxS	S	/	S	/
Ampicillin	S	PBP1a/PBP2x/ soxS	S	/	S	/
Cefuroxime	R	PBP1a/PBP2x/ soxS	R	/	S	/
Ceftiofur	S	PBP1a/PBP2x/ soxS	S	/	S	/
Cefoxitin	S	PBP1a/PBP2x/ soxS	S	/	R	/
Penicillin	R	PBP1a/PBP2x/ soxS	R	/	R	/
Gentamicin	S	gidB /baes/rpsL	S	baeS/ parY	I	baeS/ parY
Streptomycin	S	gidB /baes/rpsL	S	baeS/ parY	I	baeS/ parY
Erythromycin	S	lmrP/efrA	S	efrA/macB/oleC/oleC/mtrA/optrA/lmrC/erm(45)/ lmrP/vgaE/optrA	S	efrA/macB/oleC/oleC/mtrA/optrA/ lmrC/erm(45)/lmrP/vgaE/optrA
Doxycycline	S	lmrP/efrA	S	efrA/macB/oleC/oleC/mtrA/optrA/ lmrC/erm(45)/lmrP/vgaE/optrA	I	efrA/macB/oleC/oleC/mtrA/optrA/ lmrC/erm(45)/lmrP/vgaE/optrA
Oxytetracycline	S	rpsJ/adeR/soxS/tetT/lmrP/tetB(P)/ soxS/efrA	S	tetA(46)/tetB(46)/tetA(58)/otr(A)/optrA/tetB(P)/tetW/ lmrC/tetT/lmrP/vgaE/optrA/adeR	S	tetA(46)/tetB(46)/tetA(58)/otr(A)/optrA/tetB(P)/tetW/ lmrC/tetT/lmrP/vgaE/optrA/adeR
Tetracycline	S	rpsJ/adeR/soxS/tetT/lmrP/tetB(P)/ soxS/efrA	S	tetA(46)/tetB(46)/tetA(58)/otr(A)/optrA/ tetB(P)/tetW/lmrC/tetT/lmrP/vgaE/optrA/adeR	S	tetA(46)/tetB(46)/tetA(58)/otr(A)/optrA/tetB(P)/ tetW/lmrC/tetT/lmrP/vgaE/optrA/adeR
Levofloxacin	S	efrA/arlR/arlS/gyrA/efrB/patB/ soxS/efrA	S	efrA/arlR/arlS/arlS/patA/patB	S	efrA/arlR/arlS/arlS/patA/patB
Norfloxacin	S	efrA/arlR/arlS/gyrA/efrB/patB/ soxS/efrA	S	efrA/arlR/arlS/arlS/patA/patB	I	efrA/arlR/arlS/arlS/patA/patB
Enrofloxacin	S	efrA/arlR/arlS/gyrA/efrB/patB/ soxS/efrA	S	efrA/arlR/arlS/arlS/patA/patB	I	efrA/arlR/arlS/arlS/patA/patB
Ciprofloxacin	S	efrA/arlR/arlS/gyrA/efrB/patB/ soxS/efrA	S	efrA/arlR/arlS/arlS/patA/patB	R	efrA/arlR/arlS/arlS/patA/patB
Sulfafurazole	S	/	S	sul4	S	sul4
Sulfadiazine Sodium	S	/	S	sul4	R	sul4
Rifampin	S	rpoB/efrA/efrB/efrA/ soxS	S	rpoB2/efrA	S	rpoB2/efrA
Clindymycin	S	lmrB/cfrA/lmrP/salA	S	lmrD/cfrA/lmrC/erm(45)/lmrP/vgaE/optrA	R	lmrD/cfrA/lmrC/erm(45)/lmrP/vgaE/optrA

10.3389/fmicb.2023.1285027

10.3389/fmicb.2023.1285027

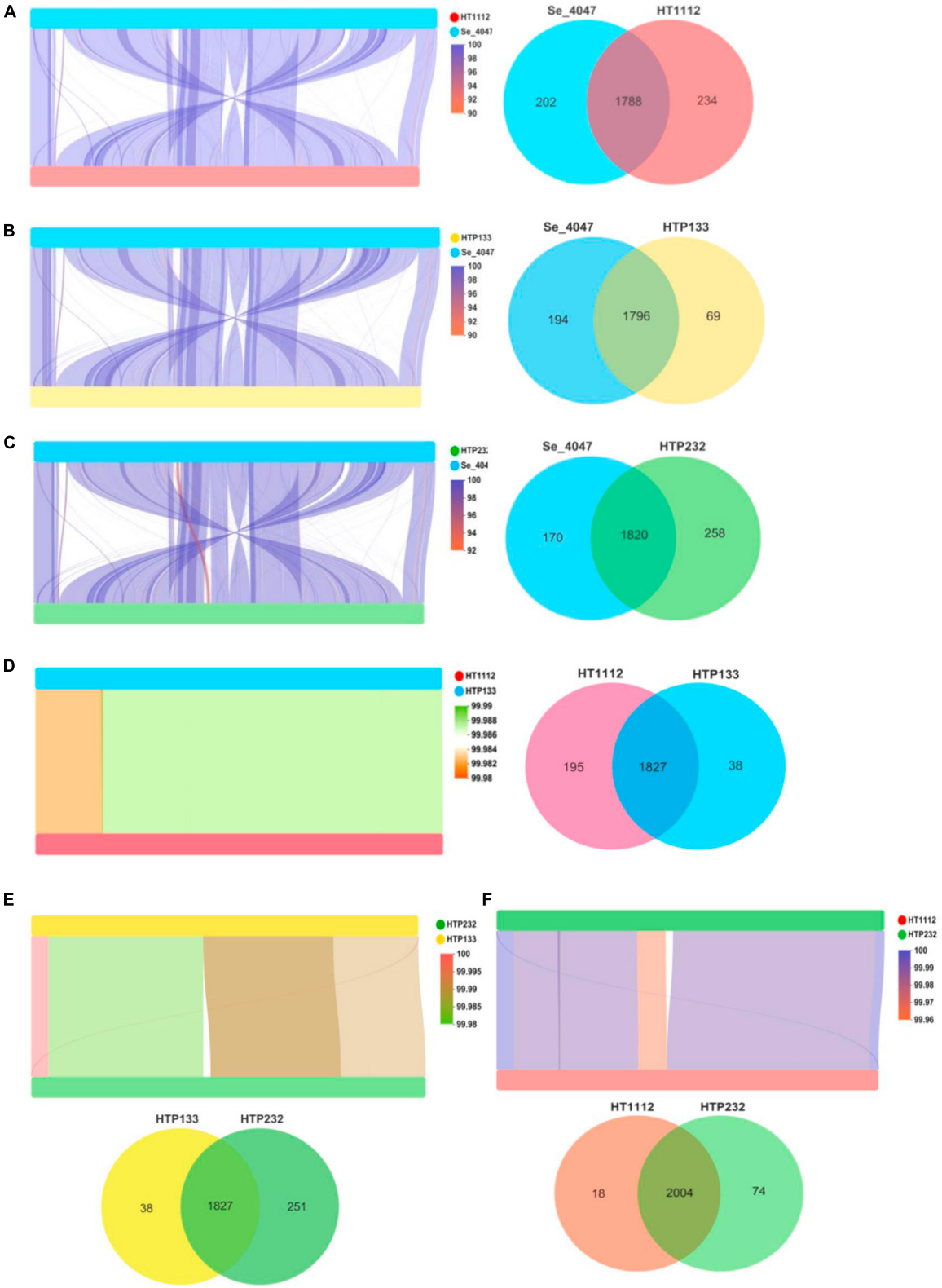


FIGURE 5
Linear comparison of *S. equi* isolate chromosomes and Venn diagrams illustrating overlap of proteins identified among **(A)** Se 4047 compared to HT1112, **(B)** Se 4047 compared to HTP133, **(C)** Se 4047 compared to HTP232, **(D)** HTP133 compared to HTP232, **(E)** HT1112 compared to HTP232, and **(F)** HT1112 compared to HTP133. Pairwise comparison of the chromosomes of **(A)** Se 4047 compared to HT1112, **(B)** Se 4047 compared to HTP133, **(C)** Se 4047 compared to HTP232, **(D)** HTP133 compared to HTP232, **(E)** HT1112 compared to HTP232, and **(F)** HT1112 compared to HTP133.

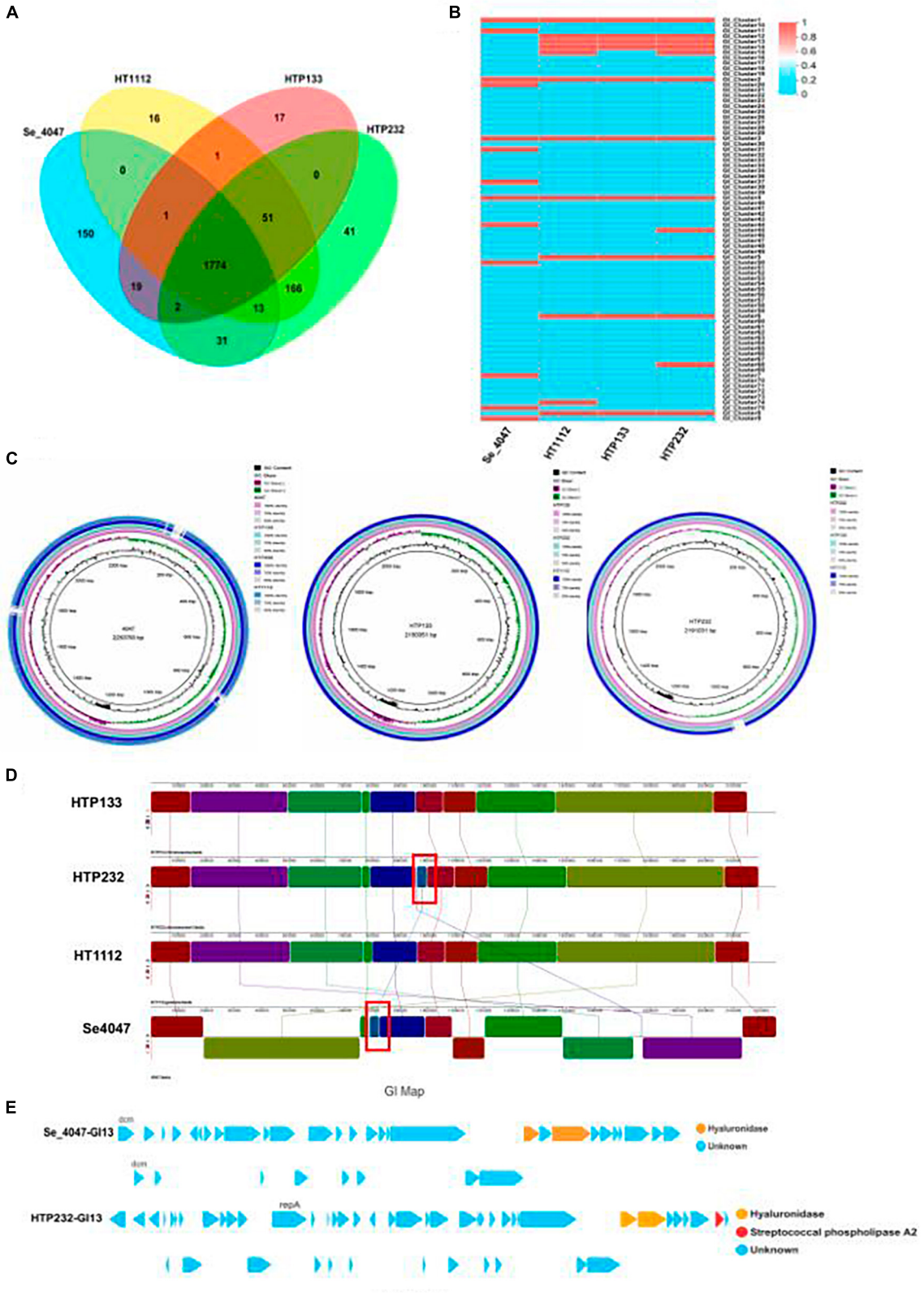


FIGURE 6
Comparison of three genomes of *S. equi* isolates to a reference strain. **(A)** The numbers of orthologous gene families and unique genes. The Venn diagram shows the number of orthologous gene families in the core genome (the center part) and the number of unique genes in each genome. The different colors indicated different sampling areas of the strains as indicated. **(B)** Heat map based on the comparison of genomic islands of three isolate genomes and a reference strain. **(C)** Alignment of *S. equi* HT1112, HTP133, and HTP232 genomes to those of reference strain Se 4047 using BRIG. **(D)** Pairwise comparison of the chromosomes of Se HTP133, Se HTP232, Se HT1112, and Se 4047 using Mauve. The image shown was generated by the Mauve rearrangement viewer. **(E)** The genomic island GI13 of Se 4047 and HTP232 harbored the hyaluronate lyase-coding gene *hya*.

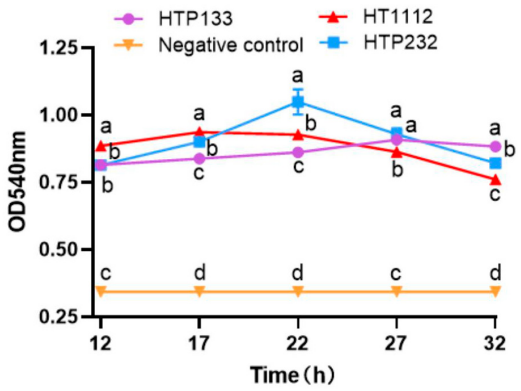


FIGURE 7
Hyaluronidase activity assay of 3 *S. equi* isolates. The enzyme activities were measured by monitoring the absorbance at 540 nm. Values represent the mean ± SD of three independent experiments. There is a significant difference ($p < 0.05$) between numbers with different letters; no significant difference ($p > 0.05$) exists between numbers with the same letter.

S. zooepidemicus has been hypothesized to have developed from an ancestor strain of *S. equi*, which possesses just two sequence types (ST), ST179 and ST151, by MLST (Webb et al., 2008). Our findings indicated that *S. equi* 4047, these three donkey stains, and both sequence types ST179.

In this study, the complete genome sequence of *S. equi* strains HT1112, HTP133, and HTP232 was determined and compared to the reference genomes of *S. equi* strain 4047. Three donkey strain genomes were found to be highly colinear with each other, but compared to the reference genomes, there is a predominant rearrangement involving a lot of inversions by pairwise and reciprocal comparisons. In addition, the HTP232 strain has more unique genes compared with *S. equi* strain 4047. Holden et al. (2009) have reported that the pathogenic specialization of *S. equi* has been shaped by a combination of gene loss and gene gain through the acquisition of mobile genetic elements by comparing the complete genomes of *S. equi* strain 4047 and *S. zooepidemicus* strain H70.

The acquisition of prophage plays an important role in the evolution of many pathogenic bacteria (Brussow et al., 2004). The prophage was absent in the SzH70 genome, and unlike SzH70, Se 4047 is polylysogenic, containing four prophages (Beres et al., 2008). Here, our result revealed that the donkey strain HTP232 contains four prophages, and they could carry two genes (*hya*) through the prophages, which may increase its survival fitness (Beres and Musser, 2007). Going deeper into virulence characterization, HTP232 contains one more virulent factor than HT1112 and HTP133.

Concerning antibiotic resistance, these three donkey strains contain the same number of ARGs; however, based on the phenotype profile, the HT1112 exhibited enhanced antimicrobial resistance. It has been reported that bacteria within biofilms have an increased resistance to antimicrobials, and infections caused by biofilm-producing bacteria are frequently resistant to antibacterial chemotherapy (Jefferson and Cerca, 2006). Other researchers also revealed that *S. zooepidemicus* can form biofilms (Yi et al., 2014). In the present study, our results with the HT1112 isolate suggest that there is a correlation between biofilm formation and their capacity for antimicrobial resistance. Grenier et al. (2009) showed that a *Streptococcus suis* serotype 2 strain isolated from a case of meningitis in pigs could form a dense biofilm and suggested a correlation between biofilm formation and the establishment of infection. Moreover, bacteria with the capacity to colonize the host by forming biofilms have more advantages in establishing persistent infections (Falkinham, 2007). Therefore, the HT1112 isolate should be a potential candidate for our future investigation against *S. equi* persistent infections.

Hyaluronic acid (HA) is a linear anionic, non-sulfated, heteropolysaccharide composed of glucuronic acid and N-acetylglucosamine joined alternately by β glycosidic bonds (Marinho et al., 2021). Most bacterial hyaluronidase is believed to facilitate the pathogen's invasion and survival in the host, which is critical to its pathogenicity and gives the pathogens an evolutionary advantage (Hynes et al., 2000; Desiere et al., 2001; Smith et al., 2005; Messina et al., 2016; Hu et al., 2021). It was reported that a few microorganisms of groups A and C *streptococci* are HA producers (Wei et al., 2012). The *S. zooepidemicus* H70 genome contains one single CDS encoding a hyaluronidase, whereas the *S. equi* 4047 genome has two hyaluronidase genes, and one hyaluronidase was encoded on a prophage (Waller and Robinson, 2013). This type of hyaluronidase was found to have a reduced substrate range and lower activity, which helps explain why *S. equi* infection rarely progresses beyond the lymphatic system (Lindsay et al., 2009). In the present study, consistent with the previous study findings in the horse isolate *S. equi* 4047, both the donkey strains HT1112 and HTP133 have one hyaluronidase encoded on a prophage. Interestingly, we note that HTP232 has two hyaluronidase genes carried on two different promoters, and one hyaluronidase gene contains a 6 bp deletion. Previous studies revealed that reduced hyaluronidase leads to high levels of hyaluronate capsules and enhances resistance to phagocytosis (Wibawan et al., 1999; Velineni and Timoney, 2015). The enhanced hyaluronate lyase activity may affect the levels of hyaluronate capsules in *S. equi* HTP232 and lead to low levels of capsules. Therefore, these results implied that the hyaluronidase gene gained by HTP232 might be critical to its pathogenicity, giving the pathogens an evolutionary advantage.

The genomic island GI13 of Se 4047 and HTP232 harbored the hyaluronate lyase-coding gene *hya*. The comparative genomic analysis of three donkey strains with Se 4047 revealed that there is horizontal gene transfer between the Se 4047 and HTP232 strains. It is worth noting that the exchange of genetic elements between host-restricted pathogens is rarely considered.

Reduced hyaluronate lyase activity provides an alternative explanation as to why *S. equi* maintains high levels of hyaluronate capsule, and in agreement with this, the Se HTP232 isolate of *S. equi* that was tested reduced virulence in a mouse challenge.

To sum up, this is the first publication that we are aware of that details the WGS and comparative analysis of the *S. equi* strain that was isolated from Chinese donkeys with strangles. The findings of our research have provided extensive insight into the phenotypic and genetic characteristics that may be used to distinguish between three isolates from donkey farms. Understanding how to prevent and cure infections may be made possible by genetic traits.

Data availability statement

Original datasets are available in a publicly accessible repository: The original contributions presented in the study are publicly available. This data can be found in the NCBI repository under accession number: PRJNA1012111.

Ethics statement

The animal study was approved by the Animal Ethics Committee, Xinjiang Agricultural University, China. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

YZ: Data curation, Investigation, Methodology, Writing – original draft. FL: Investigation, Writing – original draft. YS: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. HZ: Writing – original draft. BZ: Investigation, Writing – review & editing.

Funding

The authors 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 (grants No. U1803108).

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.

References

Alber, J., El-Sayed, A., Estoepangestie, S., Lämmler, C., and Zschöck, M. (2005). Dissemination of the superantigen encoding genes seeL, seeM, szeL and szeM in *Streptococcus equi* subsp. *equi* and *Streptococcus equi* subsp. *zooepidemicus*. *Vet. Microbiol.* 109, 135–141. doi: 10.1016/j.vetmic.2005.05.001

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 48, 517–525. doi: 10.1093/nar/gkz935

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., 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

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, 16–21. doi: 10.1093/nar/gkw387

Beres, S. B., and Musser, J. M. (2007). Contribution of exogenous genetic elements to the group A *Streptococcus* metagenome. *PLoS One* 2:e800. doi: 10.1371/journal.pone.0000800

Beres, S. B., Sesso, R., Pinto, S. W., Hoe, N. P., Porcella, S. F., Deleo, F. R., et al. (2008). Genome sequence of a lancefield group C *Streptococcus zooepidemicus* strain causing epidemic nephritis: New information about an old disease. *PLoS One* 3:e3026. doi: 10.1371/journal.pone.0003026

Bertelli, C., and Greub, G. (2013). Rapid bacterial genome sequencing:methods and applications in clinical microbiology. *Clin. Microbiol. Infect.* 19, 803–813. doi: 10.1111/1469-0691.12217

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing Group, Lau, B. Y., Hoad, G., et al. (2017). IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45, 30–35. doi: 10.1093/nar/gkx343

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., et al. (2007). CRISPR recognition tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* 8:209. doi: 10.1186/1471-2105-8-209

Boyle, A. G., Timoney, J. F., Newton, J. R., Hines, M. T., Waller, A. S., and Buchanan, B. R. (2018). *Streptococcus equi* infections in horses: Guidelines for treatment, control, and prevention of strangles-revised consensus statement. *J. Vet. Intern. Med.* 32, 633–647. doi: 10.1111/jvim.15043

Brussow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602. doi: 10.1128/MMBR.68.3.560-602.2004

Clinical and Laboratory Standards Institute [CLSI] (2020). *Performance Standards for Antimicrobial Susceptibility Testing*, 30th Edn. Wayne, PA: Clinical and Laboratory Standards Institute.

Cordoni, G., Williams, A., Durham, A., Florio, D., Zanoni, R. G., and La Ragione, R. M. (2015). Rapid diagnosis of strangles (*Streptococcus equi* subspecies *equi*) using PCR. *Res. Vet. Sci.* 102, 162–166. doi: 10.1016/j.rvsc.2015.08.008

Darling, A. E., Mau, B., and Perna, N. T. (2010). Progressive Mauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. doi: 10.1371/journal.pone.0011147

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–4641. doi: 10.1093/nar/27.23.4636

Desiere, F., McShan, W. M., van Sinderen, D., Ferretti, J. J., and Brüssow, H. (2001). Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic *Streptococci*: Evolutionary implications for prophage-host interactions. *Virology* 288, 325–341. doi: 10.1006/viro.2001.1085

Dong, N., Gao, A. S., Waller, F. R., Cook, S., Fan, D., Yuan, Y., et al. (2019). An outbreak of strangles associated with a novel genotype of *Streptococcus equi* subspecies *equi* in donkeys in China during 2018. *Equine Vet. J.* 51, 743–748. doi: 10.1111/evj.1311

Falkinham, J. O. (2007). Growth in catheter biofilms and antibiotic resistance of *Mycobacterium avium*. *J. Med. Microbiol.* 56, 250–254. doi: 10.1099/jmm.0.46935-0

Grant, J., and Stothard, P. (2008). The CGView Server: A comparative genomics tool for circular genomes. *Nucleic Acids Res.* 36, 181–184. doi: 10.1093/nar/gkn179

Grenier, D., Grignon, L., and Gottschalk, M. (2009). Characterization of biofilm formation by a *Streptococcus suis* meningitis isolate. *Vet. J.* 179, 292–295. doi: 10.1016/j.tvjl.2007.09.005

Harris, S. R., Robinson, C., Steward, K. F., Webb, K. S., Paillot, R., Parkhill, J., et al. (2015). Genome specialization and decay of the strangles pathogen, *Streptococcus equi*, is driven by persistent infection. *Genome Res.* 25, 1360–1371. doi: 10.1101/gr.189803.115

Holden, M. T., Heather, Z., Paillot, R., Steward, K. F., Webb, K., Ainslie, F., et al. (2009). Genomic evidence for the evolution of *Streptococcus equi*: Host restriction,

increased virulence, and genetic exchange with human pathogens. *PLoS Pathog.* 5:e1000346. doi: 10.1371/journal.ppat.1000346

Hu, H., Liu, H., Kweon, O., and Hart, M. E. (2021). A naturally occurring point mutation in the hyaluronidase gene (hysA1) of *Staphylococcus aureus* UAMS-1 results in reduced enzymatic activity. *Can. J. Microbiol.* 14, 1–13. doi: 10.1139/cjm-2021-0110

Hynes, W. L., Dixon, A. R., Walton, S. L., and Aridgides, L. J. (2000). The extracellular hyaluronidase gene (hylA) of *Streptococcus pyogenes*. *FEMS Microbiol. Lett.* 184, 109–112. doi: 10.1111/j.1574-6968.2000.tb08999.x

Jefferson, K. K., and Cerca, N. (2006). Bacterial-bacterial cell interactions in biofilms: Detection of polysaccharide intercellular adhesins by blotting and confocal microscopy. *Methods Mol. Biol.* 341, 119–126.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, 109–114. doi: 10.1093/nar/gkr988

Kelly, C, Bugg, M, Robinson, C, Mitchell, Z, Davis-Poynter, N, Newton, J. R., et al. (2006). Sequence variation of the SeM gene of *Streptococcus equi* allows discrimination of the source of strangles outbreaks. *J. Clin. Microbiol.* 44, 480–486. doi: 10.1128/JCM.44.2.480-486.2006

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Lindsay, A. M., Zhang, M., Mitchell, Z., Holden, M. T., Waller, A. S., Sutcliffe, I. C., et al. (2009). The *Streptococcus equi* prophage-encoded protein SEQ2045 is a hyaluronan-specific hyaluronate lyase that is produced during equine infection. *Microbiology* 155, 443–449. doi: 10.1099/mic.0.020826-0

Marinho, A., Nunes, C., and Reis, S. (2021). Hyaluronic Acid: A key ingredient in the therapy of inflammation. *Biomolecules* 11:1518. doi: 10.3390/biom11101518

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

Messina, L., Gavira, J. A., Pernagallo, S., Unciti-Broceta, J. D., Sanchez Martin, R. M., Diaz-Mochon, J. J., et al. (2016). Identification and characterization of a bacterial hyaluronidase and its production in recombinant form. *FEBS Lett.* 590, 2180–2189. doi: 10.1002/1873-3468.12258

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* 42, 206–214. doi: 10.1093/nar/gkt1226

Paillot, R., Darby, A. C., Robinson, C., Wright, N. L., Steward, K. F., Anderson, E., et al. (2010). Identification of three novel superantigen-encoding genes in *Streptococcus equi* subsp. *zooepidemicus*, szeF, szeN, and szeP. *Infect. Immun.* 78, 4817–4827. doi: 10.1128/IAI.00751-10

Rodriguez-Valera, F., Martin-Cuadrado, A. B., and López-Pérez, M. (2016). Flexible genomic islands as drivers of genome evolution. *Curr. Opin. Microbiol.* 31, 154–160. doi: 10.1016/j.mib.2016.03.014

Smith, N. L., Taylor, E. J., Lindsay, A. M., Charnock, S. J., Turkenburg, J. P., Dodson, E. J., et al. (2005). Structure of a group A *Streptococcal* phage-encoded virulence factor reveals a catalytically active triple-stranded β-helix. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17652–17657. doi: 10.1073/pnas.0504782102

Stothard, P., and Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics* 21, 537–539. doi: 10.1093/bioinformatics/bti054

Velineni, S., and Timoney, J. F. (2015). Capsular hyaluronic acid of equine isolates of *Streptococcus zooepidemicus* is upregulated at temperatures below 35°C. *Equine Vet. J.* 47, 333–338. doi: 10.1111/evj.12272

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Waller, A. S. (2013). Strangles: Taking steps towards eradication. *Vet. Microbiol.* 167, 50–60. doi: 10.1016/j.vetmic.2013.03.033

Waller, A. S., and Robinson, C. (2013). *Streptococcus zooepidemicus* and *Streptococcus equi* evolution: The role of CRISPRs. *Biochem. Soc. Trans.* 41, 1437–1443. doi: 10.1042/BST20130165

Wang, Y., Zhang, W., Wu, Z., and Lu, C. (2011). Reduced virulence is an important characteristic of biofilm infection of *Streptococcus suis*. *FEMS Microbiol. Lett.* 316, 36–43. doi: 10.1111/j.1574-6968.2010.02189.x

Webb, K., Jolley, K. A., Mitchell, Z., Robinson, C., Newton, J. R., Maiden, M. C. J., et al. (2008). Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group. *Microbiology* 154, 3016–3024. doi: 10.1099/mic.0.2008/018911-0

Wei, Z., Fu, Q., Chen, Y., Cong, P., Xiao, S., Mo, D., et al. (2012). The capsule of *Streptococcus equi* ssp. *zooepidemicus* is a target for attenuation in vaccine development. *Vaccine* 30, 4670–4675. doi: 10.1016/j.vaccine.2012.04.092

Wibawan, I. W., Pasaribu, F. H., Utama, I. H., Abdulmawjood, A., and Lämmler, C. (1999). The role of hyaluronic acid capsular material of *Streptococcus equi* subsp. *zooepidemicus* in mediating adherence to HeLa cells and in resisting phagocytosis. *Res. Vet. Sci.* 67, 131–135. doi: 10.1053/rvsc.1998.0287

Yi, L., Wang, Y., Ma, Z., Zhang, H., Li, Y., Zheng, J. X., et al. (2014). Biofilm formation of *Streptococcus equi* ssp. *zooepidemicus* and comparative proteomic analysis of biofilm and planktonic cells. *Curr. Microbiol.* 69, 227–233. doi: 10.1007/s00284-014-0574-z

Zarankiewicz, T., Madej, J., Galli, J., Bajzert, J., and Stefaniak, T. (2012). Inhibition of in vitro *Histophilus somni* biofilm production by recombinant Hsp60 antibodies. *Pol. J. Vet. Sci.* 15, 373–378. doi: 10.2478/v10181-012-0056-9

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 20 November 2023
DOI 10.3389/fmicb.2023.1275918

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Edet E. Udo,
Kuwait University, Kuwait
Miklos Fuzi,
Independent researcher, Budapest, Hungary

*CORRESPONDENCE
Agnes Marie Sá Figueiredo
✉ agnes@micro.ufrj.br

RECEIVED 10 August 2023
ACCEPTED 31 October 2023
PUBLISHED 20 November 2023

CITATION
Esteves MAC, Viana AS, Viçosa GN, Botelho AMN, Moustafa AM, Mansoldo FRP, Ferreira ALP, Vermelho AB, Ferreira-Carvalho BT, Planet PJ and Figueiredo AMS (2023) RdJ detection tests to identify a unique MRSA clone of ST105-SCC*mec*II lineage and its variants disseminated in the metropolitan region of Rio de Janeiro. *Front. Microbiol.* 14:1275918. doi: 10.3389/fmicb.2023.1275918

COPYRIGHT
© 2023 Esteves, Viana, Viçosa, Botelho, Moustafa, Mansoldo, Ferreira, Vermelho, Ferreira-Carvalho, Planet and Figueiredo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

RdJ detection tests to identify a unique MRSA clone of ST105-SCC*mec*II lineage and its variants disseminated in the metropolitan region of Rio de Janeiro

Matheus Assis Côrtes Esteves[1], Alice Slotfeldt Viana[1], Gabriela Nogueira Viçosa[1], Ana Maria Nunes Botelho[2], Ahmed M. Moustafa[3,4], Felipe Raposo Passos Mansoldo[5], Adriana Lucia Pires Ferreira[6,7], Alane Beatriz Vermelho[5], Bernadete Teixeira Ferreira-Carvalho[1], Paul Joseph Planet[3,4] and Agnes Marie Sá Figueiredo[1,8]*

[1]Departamento de Microbiologia Médica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, [2]Biomanguinhos, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [3]Children's Hospital of Philadelphia, Philadelphia, PA, United States, [4]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, United States, [5]Departamento de Microbiologia Geral, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, [6]Hospital Universitário Clementino Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, [7]Dasa Medicina Diagnóstica, Duque de Caxias, Brazil, [8]Faculdade de Medicina, Programa de Pós-graduação em Patologia, Universidade Federal Fluminense, Niterói, Brazil

Hospital bloodstream infection (BSI) caused by methicillin-resistant *Staphylococcus aureus* (MRSA) is a major cause of morbidity and mortality and is frequently related to invasive procedures and medically complex patients. An important feature of MRSA is the clonal structure of its population. Specific MRSA clones may differ in their pathogenic, epidemiological, and antimicrobial resistance profiles. Whole-genome sequencing is currently the most robust and discriminatory technique for tracking hypervirulent/well-adapted MRSA clones. However, it remains an expensive and time-consuming technique that requires specialized personnel. In this work, we describe a pangenome protocol, based on binary matrix (1,0) of open reading frames (ORFs), that can be used to quickly find diagnostic, apomorphic sequence mutations that can serve as biomarkers. We use this technique to create a diagnostic screen for MRSA isolates circulating in the Rio de Janeiro metropolitan area, the RdJ clone, which is prevalent in BSI. The method described here has 100% specificity and sensitivity, eliminating the need to use genomic sequencing for clonal identification. The protocol used is relatively simple and all the steps, formulas and commands used are described in this work, such that this strategy can also be used to identify other MRSA clones and even clones from other bacterial species.

KEYWORDS

emerging MRSA, rapid test for clonality detection, biomarker search, pangenome matrix, bacterial genomics

1 Introduction

Staphylococcus aureus is a leading cause of bloodstream infection (BSI) in hospitals and the community, which has an associated mortality rate of 20% to 30% (Kwiecinski and Horswill, 2020). The risk associated with patients over 60 years of age is even greater, with lethality reaching more than half of those infected (Loftus et al., 2022). Furthermore, BSI caused by methicillin-resistant *S. aureus* (MRSA) is recognized as an important health problem, as it appears to increase morbidity and mortality compared to those associated to methicillin-susceptible *S. aureus* (MSSA) strains (Kwiecinski and Horswill, 2020). Reported percentages of MRSA in BSI in Latin-American countries are in the alarming range of 40% to 60% of all *S. aureus* BSIs (Arias et al., 2017; Iuliano et al., 2018). An important strategy to prevent the spread of multidrug-resistant (MDR) bacteria, especially those circulating in healthcare environments, is the rapid identification and control of pathogens. High-income countries that have invested in a "search-and-destroy" policy for MDR pathogens have very low rates of MRSA in their hospitals (Souverein et al., 2016). Whole Genome Sequencing (WGS) has proven useful to accurately identify clones, track clonal transmission, and limit bacterial outbreaks caused by MDR bacteria (Yang et al., 2006; Lakhundi and Zhang, 2018; Beukers et al., 2020). The importance of identifying MRSA clones in hospitals also lies in the fact that different MRSA clones may be involved in different clinical and epidemiological scenarios and may also have different virulence and MDR profiles (Côrtes et al., 2021; Viana et al., 2021). However, WGS is a time-consuming technique that requires specialized knowledge, the use of bioinformatics programs to determine clonality with reliable power, and expensive resources that may not be available in all settings.

Recently, a unique MRSA clone, the Rio de Janeiro (RdJ) strain, of the lineage ST105 (CC5)-SCC*mec*II-002 and its variants emerged in several hospitals located in the metropolitan region of Rio de Janeiro city (Viana et al., 2021). Notably, RdJ was found more frequently in blood samples and exhibited a superior ability to evade THP-1 monocyte killing compared to its clonal complex 5 (CC5) close relatives (Viana et al., 2021). However, due to the high genetic similarity between RdJ and other CC5 strains, until now, the only approach available to accurately identify RdJ is WGS. In this work, to design a simple RdJ detection test, we created a binary matrix (1,0) that compares two groups of pangenome open reading frames (ORFs; RdJ and non-RdJ) based on presence/absence of ORFs with 100% nucleotide identity. We used this matrix to identify apomorphic mutations and find specific biomarkers to identify members of the RdJ clone.

2 Materials and methods

2.1 Development of RdJ detection tests

2.1.1 Construction of pangenome matrix

A total of 661 MRSA genomes from clonal complex 5 (CC5) were used. These genomes included MRSA strains ($N = 179$) isolated from patients hospitalized in Rio de Janeiro metropolitan region that were previously sequenced by our group (Viana et al., 2021); in addition to CC5 genomes available in the GenBank ($N = 482$). The FASTA files of these 661 CC5 genomes are accessible by the Mendeley Data Repository (doi: 10.17632/jd5tsjp4g6.1). A concatenation of all open reading frames (ORFs) from the genome sequence of strain N315 (ST5-SCC*mec*II; Acc: BA000018.3) was used as query to conduct nBlast using pangenome groups of CC5-RdJ and CC5-non-RdJ as the subject. A binary pangenomic matrix (1,0) was constructed based on presence/absence of ORFs with 100% nucleotide identity for each pangenomic group. Student's *t*-test (two tailed for independent groups) was used to calculate differences (presence/absence) for each ORF between groups. A volcano scatterplot was defined with the Log_2fold change and−Log*p* value (x,y coordinates; respectively).

The ORF sequences with the lowest *p*-values were localized in the genomes of the Brazilian strain CD16-016 (RdJ strain; Acc: GCA_021010535.1) using the "map to reference tool" (Geneious Prime software; https://www.geneious.com) and curated manually using NCBI[1] or Uniprot-SwissProt[2] databases. The Microsoft Excel[3] program formulas and commands used for these analyses are available at Mendeley Data Repository (doi: 10.17632/jd5tsjp4g6.1). Genotyping (MLST, SCC*mec* typing, and *spa* typing) was predicted for all strains analyzed from next-generation sequencing data using the Center for Genomic Epidemiology platform[4] with the software Multi Locus Sequence Typing (MLST 2.0), SCCmecFinder and spaTyper, respectively.

2.1.2 Selection of the biomarker for the development of RdJ detection tests

The distribution of the volcano plot revealed an apomorphic single nucleotide mutation (A1106G) in the *aur* gene, encoding the protease aureolysin, that differentiates members of the RdJ clade from their ancestors and close relatives. To further investigate the exclusivity of this mutation for the ST105 RdJ isolates, the coding sequences (CDS) of the total 71,017 annotated genomes of *S. aureus* were downloaded from the NCBI GenBank on April 23, 2023, using the NCBI dataset tool, and Unix xargs command was used to combine find and grep commands to search for the 20 nucleotides (TATAAGGGCCAGTCAGGCGC) in all genomes. This sequence corresponds to the location 1,099–1,118 nt of the *aur* gene, comprising the *aur* mutation (A1106G), in the genome of the RdJ strain CD16-016 (Acc: GCA_021010535.1). In addition, the sequence types (ST) of all 71,071 *S. aureus* genomes were determined using MLST tool.[5] The *aur* gene sequence of all RdJ genomes was analyzed using Geneious Prime platform for mutations that could lead to creation or deletion of endonuclease restriction sites. Analysis of the *aur* mutation (A1106G), unique to RdJ isolates, showed that it creates a new and unique restriction site for the *Bgl*I endonuclease. This feature was exploited for the design of RdJ detection tests.

2.1.3 Phylogenetic tree

A maximum likelihood tree was constructed using 180 methicillin-resistant CC5 genomes from hospitals located in the metropolitan region of Rio de Janeiro city, Brazil, in the period 2014–2017, along with 542 genomes obtained from GenBank and

1 https://blast.ncbi.nlm.nih.gov
2 https://www.uniprot.org
3 http://www.microsoft.com
4 https://www.genomicepidemiology.org/services/
5 https://github.com/tseemann/mlst

TABLE 1 Primers used for RdJ detection tests.

Primer	Direction	Sequence (5′–3′)	Amplicon (bp)	References
MecA	Forward	AGATGATACCTTCGTTCCACT	585	This study
	Reverse	CTGGTGAAGTTGTAATCTGGA		
AgrII	Forward	ATGTGTGCTCATGCAAAGTCTT	160	Beltrame et al. (2012)
	Reverse	CATGTGCATAAATAACAACGG		
Aur	Forward	ACTTGGCGATACAAAAGATATCA	752	This study
	Reverse	ATTCCGTTAATGCTCGGTAGT		

SRA (Supplementary Table S1). Single nucleotide polymorphism (SNP) alignment was performed using Realphy 1.13 pipeline.[6] The unrooted tree was built using Geneious Tree Builder[7] with Tamura-Nei distance model and Neighbor-Joining method. The tree was visualized using iTol v6.[8] We used this tree study to highlight the RdJ clade carrying the apormorphic SNP mutation (A1106G) in *aur* gene.

2.1.4 Oligonucleotide design

Our strategy was to design a method capable of (i) confirming the presence of *mecA* gene, (ii) detecting the clonal complex 5 (CC5) by the Agr type II (*agrII*; specific for CC5 strains), and (iii) identifying the RdJ isolates by detecting the *aur* mutation A1106G. Primers were designed for a multiplex PCR (mPCR) to amplify segments of *mecA*, *agr*II, and *aur* genes. Subsequently, the mPCR product was digested with the endonuclease *Bgl*I and the restriction fragment length polymorphisms (RFLP) analyzed. Table 1 shows the list of primers used for mPCR-RFLP method designed for the RdJ detection tests.

2.2 Detailed RdJ detection tests protocol

DNA was prepared using the Wizard Genomic DNA Purification Kit (Cat# A1125; Promega Corporation, Madison, WI, United States) following the manufacturer's protocol, except that a lysis step with 200 U lysostaphin (Cat# L7386; Sigma-Aldrich Brasil Ltda; São Paulo, SP, Brazil) in 500 µL of the bacterial suspension in 50 mM EDTA pH 8.0 was used to replace the enzymatic lysis suggested by the manufacturer.

The mPCR reaction was prepared using 300–400 ng DNA, 0.2 µM of each primer (forward and reverse) MecA and Aur, and 0.4 µM of AgrII primer. GoTaq Colorless Master Mix (Cat# M7133; Promega) was used as suggested by the manufacturer. Finally, the reaction was supplemented with dNTP Mix (Cat#U1515; Promega) to a final concentration of 0.4 mM in 25 µL of final volume.

The PCR program was the following: initial denaturation at 95°C for 5 min; followed by 30 cycles of denaturation at 94°C for 45 s, annealing at 52°C for 45 s and extension at 72°C for 40s; with final extension at 72°C for 7 min. Amplification was performed on a Veriti™ 96-Well ThermalCycler thermal cycler (Cat#4452299; Applied Biosystems; Foster City, CA, United States).

The mPCR product was treated with 10 U of *Bgl*I and the reaction performed as recommended by the manufacturer (Cat# R0143L; New England Biolabs, Ipswich, MA, United States). After incubation at 37°C/1 h, 1 µL of 6 X Gel Loading Dye (Cat#P1011-1; Sinapse Inc., Brazil) was added to an aliquot of 5 µL of the fragmented product and applied into the slot of a 2% agarose gel in 1 X TAE (40 mM Tris-acetate, 10 mM EDTA, pH 8.3). The DNA gel electrophoresis was performed at 120 V for 50 min. Gels were treated for 20 min with 0.5 µg/mL ethidium bromide (Cat#E7637OBS; Sigma Laboratories, Saint Louis, MO, United States) and visualized on a Gel Doc EZ System with UV light (Cat#1708270, Bio-Rad, Hercules, CA, United States).

2.3 Evaluation of RdJ detection tests

2.3.1 Bacterial collection

A collection of 217 *S. aureus* isolates was analyzed using RdJ detection tests. These isolates were tested in two steps. In the first stage, the researcher responsible for the tests previously knew the WGS data, and in the second stage, the name of each isolate was transformed into a code such that the researcher did not know which isolates were being tested (blind test). For the first stage, a total of 87 isolates were analyzed including MSSA, $N = 5$; non-CC5 MRSA, $N = 19$; ST5(CC5) MRSA, $N = 33$; ST105(CC5) RdJ, $N = 27$; ST105(CC5) non-RdJ, $N = 3$. For the second stage, a total of 130 MRSA isolates were tested including 77 non-RdJ isolates [ST5(CC5), $N = 53$; ST1635(CC5), $N = 13$; ST105(CC5), $N = 1$; and Non-CC5 MRSA, $N = 10$] and 53 RdJ-isolates [ST105(CC5), $N = 51$; and ST4876(CC5), $N = 2$]. All isolates were from our laboratory collection, and were previously genotyped based on MLST, SCC*mec* and *spa* typing using genomic approaches (Viana et al., 2021).

2.3.2 Statistical analysis

The performance of the RdJ detection tests was evaluated using the Diagnostic Test Calculator software[9] from the University of Illinois, Chicago. Data from genomic and phylogenomic strategies were used as the gold standard for comparison purpose.

6 https://realphy.unibas.ch/realphy/

7 https://www.geneious.com/features/phylogenetic-tree-building/

8 https://itol.embl.de/

9 http://araw.mede.uic.edu/cgi-bin/testcalc.pl

3 Results

3.1 Choosing a biomarker for RdJ detection tests

Using a cutoff of $p < 10^{-50}$, the results from the pangenomic matrix revealed 56 ORFs that were found predominantly among the RdJ genomes (Supplementary Table S2). The lowest *value of p* (4.81×10^{-271}) obtained in this analysis was related to the aureolysin coding sequence (*aur*; Figure 1), indicating that this ORF was a strong candidate to differentiate RdJ from other CC5 MRSA. The zinc metalloproteinase aureolysin has been associated with host immune evasion (Laarman et al., 2011). Five other genes were placed in the Volcano plot in a position close to the *aur* gene, also suggesting their potential to differentiate RdJ from other CC5 MRSA. These coding sequences were also manually analyzed and annotated using RdJ genomes, and include genes for putative products for an anthranilate synthase component, which has a role in amino acid transport/metabolism (Pohl et al., 2009); phosphoenolpyruvate carboxykinase, an enzyme essential for gluconeogenesis that plays a key role in the growth and survival of *S. aureus* cells in the absence of glucose (Harper et al., 2018); a hypothetical protein similar to OppA peptide binding protein, responsible for capturing peptides from the extracellular medium (Hiron et al., 2007); and two hypothetical proteins of unknown function (Table 2).

We analyzed the *aur* mutation (A1106G) to look for possible alteration in an endonuclease site that would allow a simple method for detecting this mutation and found that the A1106G created a *BglI* site in RdJ genomes (Figure 2). Search analysis using the 20 nucleotides around the A1106G mutation as a reference against 71,017 *S. aureus* genomes showed that all RdJ genomes (the ones grouped in the RdJ clade; colored in yellow) carried this mutation (Figure 3). Manual curation of the entire *aur* sequence of RdJ genomes showed that only the genomes of strains CD15-152, CD16-151, CD16-152, and UB641 had <100% nucleotide identity (Supplementary Table S1). Each of these strains carries another single-nucleotide mutation, in addition to the expected A1106G, which does not create a *BglI* site, and should not interfere with the detection of RdJ isolates.

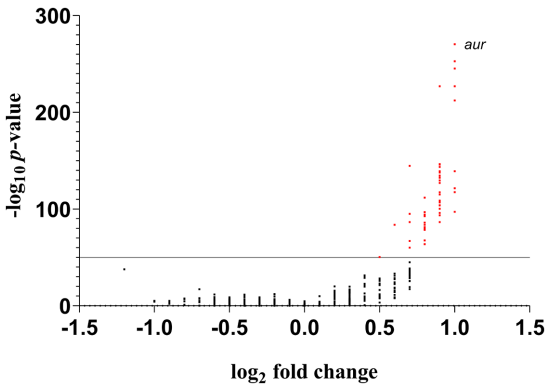


FIGURE 1
Volcano plot of the distribution of ORFs that differ between the RdJ and all other CC5 MRSA genomes. The red dot represents ORFs whose difference in distributions resulted in *p*-values $\leq 10^{-50}$.

The analysis of 71,017 *S. aureus* genomes from NCBI database (Supplementary Table S3) revealed a total of 1,385 ST105 sequenced genomes. One hundred and sixty of 71,017 have A1106G mutation in the *aur* gene (Supplementary Table S4). From the 160 genomes, 136 were classified as ST105, 14 show ambiguous results but were considered ST105 by manual curation, and 4 are ST105 variants [ST4876 ($n = 2$) and ST5288 ($n = 2$)] that cluster within RdJ clade ($n = 154$). The remaining six genomes are from different CCs: ST8 (0.02%; $n = 2/11,198$), ST12 (0.79%; $n = 2/253$), ST30 (0.03%; $n = 1/3,015$) and ST479 (4.55%; $n = 1/22$). In the RdJ clade we added six more ST105 RdJ genomes from Rio de Janeiro that could not be downloaded together with the 71,017 *S. aureus* genomes, making a total of 160 RdJ genomes (Supplementary Table S1).

BioSample analysis for the 154 genomes classified as RdJ using NCBI entrez-direct revealed that all were from Brazil, except two from the Netherlands (Supplementary Tables S1, S4). Of the 152 Brazilian genomes, 79 were from Rio de Janeiro and were previously sequenced by our group (Viana et al., 2021). The remaining 73 Brazilian genomes were mostly submitted by the Universidad El Bosque, Colombia and were from MRSA isolates collected from blood in hospitals in the city of São Paulo, where the RdJ clone was predominant among MRSA isolates (53.97%; $n = 68/126$; Arias et al., 2017; Supplementary Table S4). These results confirm that the vast majority of ST105 isolates detected in Brazil (RdJ clone) evolved to acquire a specific *aur* mutation that was conserved during the rapid expansion of this clone in this country.

3.2 RdJ detection tests

Based on the phylogenomic findings, we designed the multiplex PCR RdJ detection test protocol that is described in detail in the Material & Methods section. This test allows detection of *mecA*, presence/absence of CC5 (predominant clonal complex in Rio de Janeiro), and differentiation of CC5-RdJ from CC5-non-RdJ MRSA isolates (Figure 4).

Once the RdJ detection tests were established, a total of 87 *Staphylococcus* isolates were initially tested. These results are compiled in Supplementary Table S5. All mPCR patterns obtained with RdJ detection tests matched to the expected genotyping and phylogenetic profile found through genomic analyses. After this step, we tested another 130 MRSA isolates in a blind test. Again, 100% of congruence was observed between the results of the RdJ detection tests and genomic phylogeny and genotypes (Supplementary Table S5).

Although the RdJ detection tests were 100% sensitive and 100% specific for the detection of MRSA isolates of the RdJ clone (which mostly belong to the ST105-SCC*mec*II lineage, with exception of two ST4876-SCC*mec*II-t002 isolates), it is important to emphasize that these tests do not have the same sensitivity and specificity for the identification of ST105-SCC*mec*II isolates from Rio de Janeiro that are not in the RdJ clade. Even though the vast majority of RdJ isolates from Rio de Janeiro belong to the ST105-SCC*mec*II lineage, four ST105-SCC*mec*II isolates from Rio de Janeiro (CD15-245, CD15-276, CR15-001, CR14-025) are non-RdJ clade isolates and do not have the A1106G mutation. As expected, these four strains were phylogenetically distinct from the RdJ clone isolates, grouping in another clade along with genomes of isolates from São Paulo, Brazil; Boston, USA; and Venezuela. Furthermore, in the international context, according to our BLAST analysis, non-RdJ ST105-SCC*mec*II

TABLE 2 Annotation of the six ORFs that showed the highest degree of significance in relation to the ORF differences found between the two CC5 MRSA groups of pangenomes: RdJ and non-RdJ.

ORF product	log2 *fold-change*	−log10 *p*-value	Predicted product	Mutation (nucleotide)	Type of mutation
Aur	1	270.3	Zinc metalloproteinase aureolysin	A1106G	Nonsynonymous
Trpg	1	252.7	Anthranilate synthase component	A337G	Nonsynonymous
Pcka	1	245.3	Phosphoenolpyruvate carboxykinase	A406G	Synonymous
Sa0849	0.9	226.9	Hypothetical protein, similar to peptide binding protein OppA	C1597T	Nonsynonymous
Sa1437	1	226.9	Hypothetical protein	T325C	Synonymous
Sa0667	1	212.3	Hypothetical protein	A406G	Synonymous

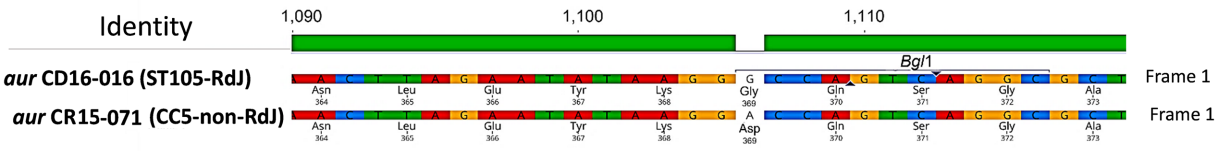


FIGURE 2
Alignment of nucleotide sequences corresponding to segments of the alleles found in CC5 strains RdJ and non-RdJ comprising the A1106G mutation region. The genomes of MRSA strains CD16-016 (ST105-RdJ, Acc: JAECLS000000000) and MRSA CR15-071 (ST5-SCC*mec*II-t539, non-RdJ, Acc: CP065868) were used to obtain this alignment on Geneious Prime pipeline.

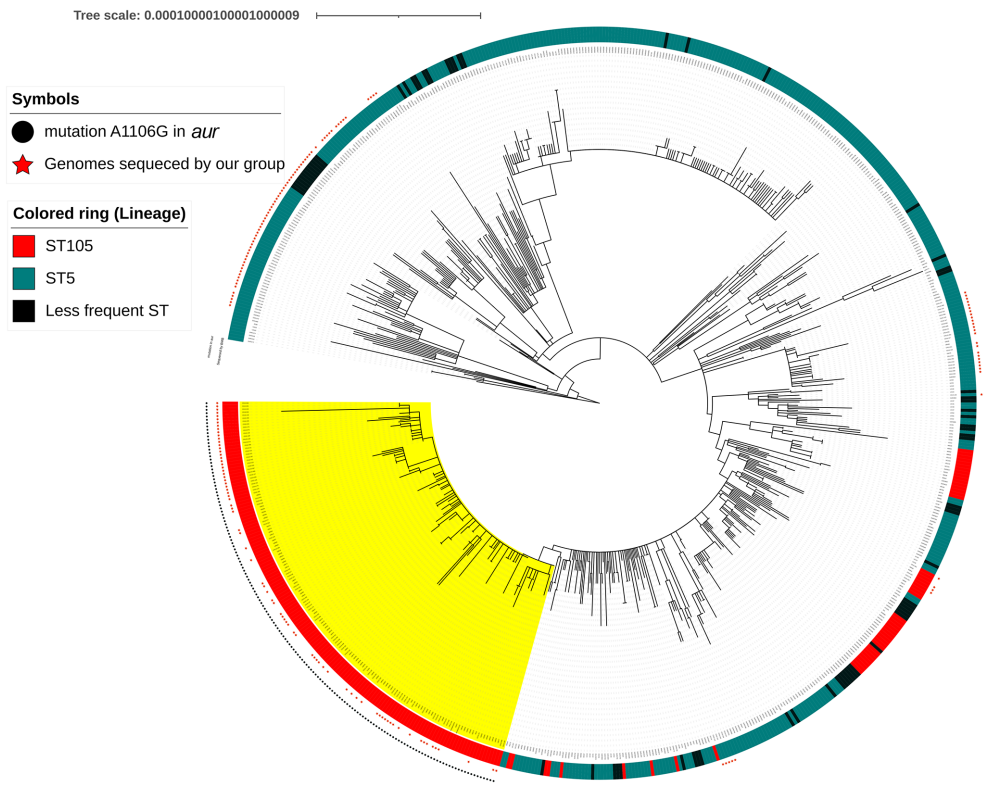


FIGURE 3
Maximum likelihood phylogenetic tree composed of 180 CC5 genomes from the metropolitan area Rio de Janeiro, sequenced by our group (red star) and 548 CC5 genomes from NCBI databases. The RdJ clade (marked with yellow color) is composed of the ST105(CC5)-SCC*mec*II isolates and few variants: ST4876(CC5)-SCC*mec*II ($n = 2$) and ST5288(CC5)-SCC*mec*II ($n = 2$). Black circle represents the apomorphic mutation (A1106G) in the *aur* gene of the RdJ genomes. The scale indicates substitutions by site.

10.3389/fmicb.2023.1275918

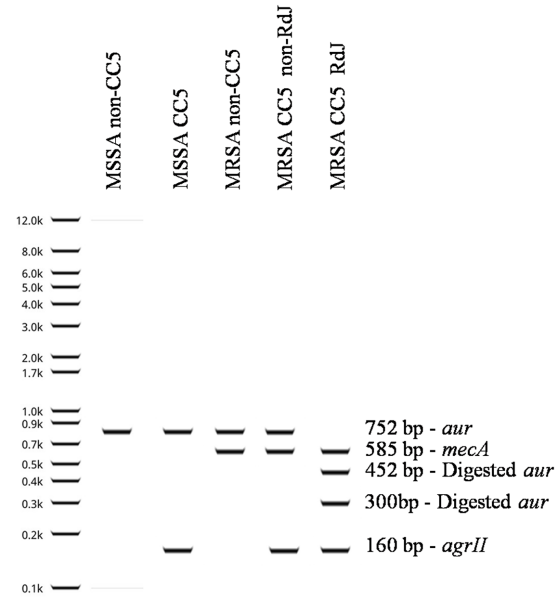


FIGURE 4
Representation of the RdJ detection tests performed for different types of *Staphylococcus aureus* strains. Shown in this figure are the possible results of the RdJ tests: MSSA non-CC5, MSSA CC5, MRSA non-CC5, MRSA CC5 non-RdJ and MRSA CC5 RdJ. Note that the MRSA CC5 RdJ isolates have two *aur* fragments at 452 bp and 300 bp, whereas MRSA CC5-non-RdJ have wild-type *aur* that lacks a *Bgl*I restriction site. MRSA is distinguished from MSSA by the amplification of a fragment of the *mecA* gene, and CC5 isolates are distinguished from non-CC5 by the *agr*II fragment.

do not have the A1106G mutation. The sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) of the RdJ detection tests in relation to their power to identify the ST105-SCC*mec*II lineage from the collection of CC5 MRSA from Rio de Janeiro studied here are presented in Table 3.

TABLE 3 Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for RdJ detection tests for the identification of ST105-SCC*mec*II (RdJ and non-RdJ) in the collection of *Staphylococcus aureus* studied.

Genotype	Detection of ST105-SCC*mec*II		Total
	Negative	Positive	
ST105-RdJ	0	78 (true)	78
ST105-non RdJ	4 (false)	0	4
ST4876	0	2 (false)	2
Other STs	128 (true)	0	128
Total	132	80	212
Specificity:	98.46% (94.55–98.81%)	PPV:	97.50% (90.78–93.65%)
Sensitivity:	95.12% (87.98–98.66%)	NPV:	96.97% (92.48–98.81%)

MLST and clone type (RdJ or non-RdJ) were performed using genomics and phylogenomics analyses. The parentheses correspond to the 95.00% confidence interval.

4 Discussion

We used a strategy of producing a binary (1,0) pangenome matrix based on the presence/absence of exact match ORFs to identify nucleotide apomorphies for rapid and simple tests for the detection of isolates belonging to the RdJ clone of MRSA. We demonstrated that the use of pangenomes is particularly powerful for identifying highly specific diagnostic tests for clonal identification. The tests developed here showed 100% specificity and sensitivity for the detection of MRSA isolates belonging to RdJ clone. Furthermore, the RdJ detection tests were equally accurate in identifying clonal complex 5 (CC5) isolates, the main clonal complex spread in Rio de Janeiro, which is also widespread in other countries including USA (Viana et al., 2021). Thus, because these tests are relatively simple and inexpensive compared with WGS, the RdJ detection tests can be used, for example, on a large scale to investigate the prevalence and incidence of the RdJ clone in hospital BSI, allowing comprehensive studies on the risks for its acquisition, and for morbidity and mortality associated with BSI caused by RdJ isolates. It is also notable that ST105-SCC*mec*II isolates increasingly have been reported more recently in hospitals in other Brazilian states and in other countries (Dabul and Camargo, 2014; Bes et al., 2021). Due to its prevalence in BSI in our region, the use of large-scale molecular tests to screen for RdJ is of great interest.

A search of the 71,017 annotated genomes deposited in NCBI GenBank database revealed that the vast majority of A1106G mutation in the *aur* gene (96.25%; $n = 154/160$) was in ST105 genomes or its variants. Although this specific mutation in non-CC5 genomes is rare, we have included specific primers in the RdJ detection tests to identify CC5. Thus, the presence of A1106G mutation in non-CC5 isolates will be revealed by the tests developed here, avoiding confusion in the identification of the RdJ clone, since the latter belongs to CC5. None of 9,555 ST5(CC5) genomes (close relatives of RdJ strains) carry A1106G mutation.

To find a biomarker to detect the RdJ clone, we initially established an arbitrary cutoff corresponding to a p-value $< 10^{-50}$. However, if none of the 56 ORFs selected by this cutoff served as a good biomarker for designing a relatively simple molecular test, a higher cutoff could be attempted, but this could compromise the specificity of the biomarker. In this case, an alternative could be to use more than one biomarker to guarantee the specificity of the detection test. Nevertheless, the need to choose a not very specific biomarker seems unlikely, given the relatively high number of commercially available restriction endonucleases. For example, in the case of the RdJ detection tests, the ORF that presented the lowest p value (*aur*) proved to be ideal for detecting the RdJ clone.

Although most isolates from Rio de Janeiro and São Paulo belong to the RdJ clone, which is part of the ST105-SCC*mec*II lineage, this method was not designed and cannot be used for identification of other isolates of ST105-SCC*mec*II lineage that do not belong to the RdJ clone. For example, out of a total 1,385 ST105 in the collection of 71,017 *S. aureus* genomes, only 150 (10.83%) had the *aur* mutation. In addition, four ST105 variants [ST4876 (n = 2) and ST5288 (n = 2)] also clustered in RdJ clade. In fact, most ST105-SCC*mec*II isolates of international origin do not contain the *aur* (A1106G) mutation. Of

the total of 71,017 *S. aureus* genomes analyzed, 5 were from Europe (Denmark, $n = 3$; Netherlands, $n = 2$) and 2 USA (city of Boston, $n = 1$ and New York City, $n = 1$).

In this context, research associating phylogenomics, genomic epidemiology and pangenomics tools is of paramount importance, not only to accelerate the understanding of evolutionary events and epidemiological factors that may be involved in the spread of hypervirulent and multidrug resistant pathogens, but also to provide simpler and economical tools (Uhlemann et al., 2014; Planet, 2017; Mirande et al., 2018). Such rapid tests can also be fundamental for designing protocols for more personalized and efficient treatment of patients, contributing to patient health safety (Rajapaksha et al., 2019).

Pickens et al. (2021) reported a dramatic decrease in the use of vancomycin as the drug of choice for the treatment of pneumonia in hospitals after the implementation of a simple PCR-based screening method to exclude the presence of MRSA. Considering that specific clones of MRSA, not infrequently, present different profiles of antimicrobial resistance, and that they may be involved in different epidemiologic and clinical scenarios and presentations (Côrtes et al., 2021; Viana et al., 2021), the use of rapid tests to identify hypervirulent, highly transmissible, well-adapted, or MDR clones is highly desirable.

Indeed, rapid molecular tests have been developed for the detection of MRSA lineages. A duplex PCR-based strategy was used to detect the livestock-associated MRSA (LA-MRSA) CC398 MRSA (Stegger et al., 2011). Also, a rapid test for the detection of the USA300 MRSA was developed resulting in a 100% match with the WGS data (Bonnstetter et al., 2007) This clone is an important, disseminated MRSA in hospital and community settings in the US (Planet, 2017). Other PCR-based tests have also been developed to identify MLST clonal complexes (Lindsay and Sung, 2010). The difference between the tests developed here and others is that the RdJ detection tests identify the clone, according to its phylogeny, while the others identify at most the MLST clonal complex.

In addition to the importance of RdJ detection tests for large-scale clinical and epidemiological investigations, we believe that scientists guided by the strategy used in this work will be able to easily perform analysis of pangenome ORFs to reveal apomorphies that can be used for the development of accurate diagnostic tests, not only for MRSA clones, but also for other pathogens, aiming at more effective treatment and successful control of infectious diseases.

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

ME: Methodology, Validation, Data curation, Formal Analysis, Investigation, Writing – original draft. AVi: Data curation, Formal Analysis, Investigation, Methodology, Writing – original draft. GV: Formal Analysis, Investigation, Writing – original draft. AB: Formal Analysis, Investigation, Writing – original draft. AM: Formal Analysis, Investigation, Writing – original draft. FM: Formal Analysis, Investigation, Writing – original draft. AFe: Formal Analysis, Investigation, Writing – original draft. AVe: Formal Analysis, Investigation, Writing – original draft. BF: Data curation, Project administration, Supervision, Writing – review & editing. PP: Data curation, Writing – review & editing, Formal Analysis, Funding acquisition, Investigation. AFi: Funding acquisition, Writing – review & editing, Conceptualization, Methodology, Project administration, Supervision, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), #307672/2019 to AFi; Fundação Carlos Chagas Filho de Apoio à Ciência (FAPERJ), #E-26/210.875/2016, E-26/210.110/2018, E-26/211.554/2019, and E-26/200.952/2021 to AFi; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), #001 to AFi; CNPq/MCTI/CT-Saúde # 408725/2022-2 to AFi; The Brazilian Ministry of Health/DECIT/CNPq, #443804/2018-4 to AFi; Bill & Melinda Gates Foundation, #INV-007641 to AFi—under the grant condition of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the author accepted manuscript version that might arise from this submission; and 1R01AI137526-01 to PP. AVi and GV received fellowships from CNPq (DTI-A) #381239/2023-3 and (PDJ) #156537/2018-4.

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.2023.1275918/full#supplementary-material

References

Arias, C. A., Reyes, J., Carvajal, L. P., Rincon, S., Diaz, L., Panesso, D., et al. (2017). A prospective cohort multicenter study of molecular epidemiology and phylogenomics of *Staphylococcus aureus* bacteremia in nine Latin American countries. *Antimicrob. Agents Chemother.* 61, e00816–e00817. doi: 10.1128/aac.00816-17

Beltrame, C. O., Botelho, A. M. N., Silva-Carvalho, M. C., Souza, R. R., Bonelli, R. R., Ramundo, M. S., et al. (2012). Restriction modification (RM) tests associated to additional molecular markers for screening prevalent MRSA clones in Brazil. *Eur. J. Clin. Microbiol. Infect. Dis.* 31, 2011–2016. doi: 10.1007/s10096-011-1534-1

Bes, T. M., Perdigão-Neto, L., Martins, R. R., Heijden, I., Trindade, P. D. A., Camilo, G., et al. (2021). Susceptibility to chlorhexidine and mupirocin among methicillin-resistant *Staphylococcus aureus* clinical isolates from a teaching hospital. *Rev. Inst. Med. Trop. São Paulo*. 63:e27. doi: 10.1590/S1678-9946202163027

Beukers, A. G., Newton, P., Hudson, B., Ross, K., Gottlieb, T., O'Sullivan, M., et al. (2020). A multicentre outbreak of ST45 MRSA containing deletions in the *spa* gene in New South Wales, Australia. *J. Antimicrob. Chemother.* 75, 1112–1116. doi: 10.1093/jac/dkz560

Bonnstetter, K. K., Wolter, D. J., Tenover, F. C., McDougal, L. K., and Goering, R. V. (2007). Rapid multiplex PCR assay for identification of USA300 community-associated methicillin-resistant *Staphylococcus aureus* isolates. *J. Clin. Microbiol.* 45, 141–146. doi: 10.1128/JCM.01228-06

Côrtes, M. F., Botelho, A. M. N., Bandeira, P. T., Mouton, W., Badiou, C., Bes, M., et al. (2021). Reductive evolution of virulence repertoire to drive the divergence between community- and hospital-associated methicillin-resistant *Staphylococcus aureus* of the ST1 lineage. *Virulence* 12, 951–967. doi: 10.1080/21505594.2021.1899616

Dabul, A. N., and Camargo, I. L. (2014). Molecular characterization of methicillin-resistant *Staphylococcus aureus* resistant to tigecycline and daptomycin isolated in a hospital in Brazil. *Epidemiol. Infect.* 142, 479–483. doi: 10.1017/S0950268813001325

Harper, L., Balasubramanian, D., Ohneck, E. A., Sause, W. E., Chapman, J., and Mejia-Sosa, B. (2018). *Staphylococcus aureus* responds to the central metabolite pyruvate to regulate virulence. *MBio* 9, e02272–e02217. doi: 10.1128/mBio.02272-17

Hiron, A., Borezée-Durant, E., Piard, J. C., and Juillard, V. (2007). Only one of four oligopeptide transport systems mediates nitrogen nutrition in *Staphylococcus aureus*. *J. Bacteriol.* 189, 5119–5129. doi: 10.1128/JB.00274-07

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., et al. (2018). Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet* 391, 1285–1300. doi: 10.1016/S0140-6736(17)33293-2

Kwiecinski, J. M., and Horswill, A. R. (2020). *Staphylococcus aureus* bloodstream infections: pathogenesis and regulatory mechanisms. *Curr. Opin. Microbiol.* 53, 51–60. doi: 10.1016/j.mib.2020.02.005

Laarman, A. J., Ruyken, M., Malone, C. L., van Strijp, J. A., Horswill, A. R., and Rooijakkers, S. H. (2011). *Staphylococcus aureus* metalloprotease aureolysin cleaves complement C3 to mediate immune evasion. *J. Immunol.* 186, 6445–6453. doi: 10.4049/jimmunol.1002948

Lakhundi, S., and Zhang, K. (2018). Methicillin-resistant *Staphylococcus aureus*: molecular characterization, evolution, and epidemiology. *Clin. Microbiol.* 31, e00020–e00018. doi: 10.1128/CMR.00020-18

Lindsay, J. A., and Sung, J. M. (2010). The RM test for determining methicillin-resistant *Staphylococcus aureus* lineages. *Methods Mol. Biol.* 642, 3–11. doi: 10.1007/978-1-60327-279-7_1

Loftus, M. J., Young-Sharma, T. E., Wati, S., Badoordeen, G. Z., Blakeway, L. V., Byers, S. M., et al. (2022). Epidemiology, antimicrobial resistance and outcomes of *Staphylococcus aureus* bacteraemia in a tertiary hospital in Fiji: a prospective cohort study. *Lancet. Reg. Health West. Pac.* 22:100438. doi: 10.1016/j.lanwpc.2022.100438

Mirande, C., Bizine, I., Giannetti, A., Picot, N., and van Belkum, A. (2018). Epidemiological aspects of healthcare-associated infections and microbial genomics. *Eur. J. Clin. Microbiol. Infect. Dis.* 37, 823–831. doi: 10.1007/s10096-017-3170-x

Pickens, C. I., Qi, C., Postelnick, M., Paonessa, J., Donnelly, H. K., and Wunderink, R. G. (2021). Association between a rapid diagnostic test to detect methicillin-resistant *Staphylococcus aureus* pneumonia and decreased vancomycin use in a medical intensive care unit over a 30-month period. *Infect. Control Hosp. Epidemiol.* 42, 1385–1387. doi: 10.1017/ice.2020.1406

Planet, P. J. (2017). Life after USA300: the rise and fall of a superbug. *J Infect Dis* 215, S71–S77. doi: 10.1093/infdis/jiw444

Pohl, K., Francois, P., Stenz, L., Schlink, F., Geiger, T., Herbert, S., et al. (2009). CodY in *Staphylococcus aureus*: a regulatory link between metabolism and virulence gene expression. *J. Bacteriol.* 191, 2953–2963. doi: 10.1128/JB.01492-08

Rajapaksha, P., Elbourne, A., Gangadoo, S., Brown, R., Cozzolino, D., and Chapman, J. (2019). A review of methods for the detection of pathogenic microorganisms. *Analyst* 144, 396–411. doi: 10.1039/c8an01488d

Souverein, D., Houtman, P., Euser, S. M., Herpers, B. L., Kluytmans, J., and Den Boer, J. W. (2016). Costs and benefits associated with the MRSA search and destroy policy in a hospital in the region Kennemerland, The Netherlands. *PloS One* 11:e0148175. doi: 10.1371/journal.pone.0148175

Stegger, M., Lindsay, J. A., Moodley, A., Skov, R., Broens, E. M., and Guardabassi, L. (2011). Rapid PCR detection of *Staphylococcus aureus* clonal complex 398 by targeting the restriction-modification system carrying *sau1-hsdS1*. *J. Clin. Microbiol.* 49, 732–734. doi: 10.1128/JCM.01970-10

Uhlemann, A. C., Otto, M., Lowy, F. D., and DeLeo, F. R. (2014). Evolution of community- and healthcare-associated methicillin-resistant *Staphylococcus aureus*. *Infect. Genet. Evol.* 21, 563–574. doi: 10.1016/j.meegid.2013.04.030

Viana, A. S., Botelho, A. M. N., Moustafa, A. M., Boge, C. L., Ferreira, A. L. P., da Silva Carvalho, M. C., et al. (2021). Multidrug-resistant methicillin-resistant *Staphylococcus aureus* associated with bacteremia and monocyte evasion, Rio de Janeiro, Brazil. *Emerg. Infect. Dis.* 27, 2825–2835. doi: 10.3201/eid2711.210097

Yang, J. A., Park, D. W., Sohn, J. W., and Kim, M. J. (2006). Novel PCR-restriction fragment length polymorphism analysis for rapid typing of staphylococcal cassette chromosome *mec* elements. *J. Clin. Microbiol.* 44, 236–238. doi: 10.1128/JCM.44.1.236-238.2006

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 07 December 2023
DOI 10.3389/fmicb.2023.1264939

Check for updates

OPEN ACCESS

EDITED BY
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

REVIEWED BY
Rosa Estela Quiroz Castañeda,
Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP), Mexico
Jianmin Zhong,
Cal Poly Humboldt, United States

*CORRESPONDENCE
Zhen-Hua Lu
✉ lucky18710360310@163.com
Zhong-Jun Shao
✉ 13759981783@163.com

†Author note

RECEIVED 21 July 2023
ACCEPTED 12 September 2023
PUBLISHED 07 December 2023

CITATION
Zhang L, Han J, Zhou Q, He Z, Sun S-W, Li R, Li R-S, Zhang W-K, Wang Y-H, Xu L-L, Lu Z-H and Shao Z-J (2023) Differential microbial composition in parasitic vs. questing ticks based on 16S next-generation sequencing. *Front. Microbiol.* 14:1264939. doi: 10.3389/fmicb.2023.1264939

COPYRIGHT
© 2023 Zhang, Han, Zhou, He, Sun, Li, Li, Zhang, Wang, Xu, Lu and Shao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Differential microbial composition in parasitic vs. questing ticks based on 16S next-generation sequencing

Lin Zhang†, Jing Han†, Qi Zhou†, Zhen He, Shi-Wei Sun, Rui Li, Rui-Shan Li, Wen-Kai Zhang, Yu-Hua Wang, Lin-Li Xu, Zhen-Hua Lu* and Zhong-Jun Shao*

Department of Epidemiology, Ministry of Education Key Lab of Hazard Assessment and Control in Special Operational Environment, School of Public Health, Air Force Medical University, Xi'an, Shaanxi, China

Introduction: As tick-borne diseases rise to become the second most prevalent arthropod-transmitted disease globally, the increasing investigations focus on ticks correspondingly. Factors contributed to this increase include anthropogenic influences, changes in vertebrate faunal composition, social-recreational shifts, and climatic variation. Employing the 16S gene sequence method in next-generation sequencing (NGS) allows comprehensive pathogen identification in samples, facilitating the development of refined approaches to tick research omnidirectionally.

Methods: In our survey, we compared the microbial richness and biological diversity of ticks in Wuwei City, Gansu province, differentiating between questing ticks found in grass and parasitic ticks collected from sheep based on 16S NGS method.

Results: The results show *Rickettsia*, *Coxiella*, and *Francisella* were detected in all 50 *Dermacentor nuttalli* samples, suggesting that the co-infection may be linked to specific symbiotic bacteria in ticks. Our findings reveal significant differences in the composition and diversity of microorganisms, with the *Friedmanniella* and *Bordetella* genera existing more prevalent in parasitic ticks than in questing ticks ($p < 0.05$). Additionally, the network analysis demonstrates that the interactions among bacterial genera can be either promotive or inhibitive in ticks exhibiting different lifestyles with the correlation index $|r| > 0.6$. For instance, *Francisella* restrains the development of 10 other bacteria in parasitic ticks, whereas *Phyllobacterium* and *Arthrobacter* enhance colonization across all tick species.

Discussion: By leveraging NGS techniques, our study reveals a high degree of species and phylogenetic diversity within the tick microbiome. It further highlights the potential to investigate the interplay between bacterial genera in both parasitic and questing ticks residing in identical habitat environments.

KEYWORDS

parasitic tick, questing tick, symbiotic bacteria, microbiome analysis, network diagram

1 Introduction

Tick-borne diseases remain a longstanding and burgeoning global health concern, with disease incidences seeing a dramatic increase since the 20th century (Rodino et al., 2020). This trend is evident worldwide and reflected in the rising number of tick-borne encephalitis and hemorrhagic fever cases reported across Europe and Asia (Rochlin and Toledo, 2020). In China,

for instance, tick-borne diseases attributed to over 40 diverse pathogens have proliferated across several provinces, including Anhui, Henan, Inner Mongolia, Tibet, Zhejiang, and Liaoning, since initial reporting in 1982 with the amplification of epidemic area and infected human amount (Lan-Hua and Yi, 2019). Notwithstanding these broad findings, studies specific to Gansu Province remain scarce. However, the importance of investigating the intricate role of microbiology in environmental contexts, given ticks' critical function as bacterial vectors, is well-documented in other regions of China.

A deeper understanding of the composition and diversity of microbiomes influencing pathogen transmission necessitates a nuanced exploration of both parasitic (feeding) ticks and questing (off-host) ticks. Differentiating these two tick lifestyles may yield crucial insights into disease manifestation in relation to varying host or off-host circumstances (Ciebiera et al., 2021), although the interplay between tick communities and the pathogens they carry remains somewhat nebulous (Kahl, 2018). Although parasitic ticks predominantly inhabit the body surfaces of hosts such as humans and domestic animals, facilitating pathogen transmission, the comparative studies of these ticks with their counterparts living in wild grass and scrub are hampered by unilateral approaches that focus narrowly on specific tick categories and pathogen transmission routes.

In the last 20 years, various techniques have been employed to elucidate the microbial composition within tick communities. The advent of next-generation sequencing (NGS) has accelerated pathogen identification, providing a rapid and efficient means to comprehensively explore pathogen community structure (Sanschagrin and Yergeau, 2014). Utilized extensively in bacterial gene detection, NGS offers high efficiency and accuracy in identifying microbes from large sample sets in a short time frame (Chaorattanakawee et al., 2022). The conservation and variability within the small subunit ribosomal RNA gene, especially the 16S rRNA gene, permit the design of unique primers for specific bacterial species, making NGS a powerful tool for disease surveillance and comprehensive microbial taxonomic identification (Rodino et al., 2021).

For this study, we selected Wuwei City in Gansu province, China, known for its rich species diversity, as our sampling site (Sun et al., 2016). Previous reports have identified new species like *Ornithodoros huajianensis* in the city's Mongolian marmots (Sun et al., 2019), as well as the presence of *Babesia* spp. and *Theileria* spp. in cattle (Sun et al., 2020). Our research aims to elucidate the differences in microbiomes and tick-borne pathogens in Wuwei City, Gansu Province. A thorough investigation of both questing and parasitic ticks will be conducted, intending to bridge the gap in current studies that overlook the microbiome composition and bacterial transmission differences between these two lifestyle ticks in natural environments. Ultimately, we intend to enhance understanding of tick-borne pathogens by detailing every procedure of this study, from tick collection to NGS analysis, with a focus on characterizing the questing ticks and the feeding ticks from sheep in Wuwei City.

2 Materials and methods

2.1 Ticks collection and morphological observation

All ticks were collected from Wuwei City (102.65°E, 37.94°N), Gansu province, in May 2022. Feeding ticks were carefully extracted from sheep using tweezers while questing ticks were gathered from grassland using the Flag Cloth Law method, which involves dragging a 90 cm × 60 cm white flannelette flag across the surface. It should be noted that the sampling of both questing and feeding ticks was opportunistic and was therefore not standardized with respect to collection area or time.

Following collection, all ticks were immediately stored in tubes with perforated lids and refrigerated at 4°C to minimize tick mortality. Afterward, ticks' surfaces were cleaned using 75% ethyl alcohol, and identification was carried out with a stereoscopic microscope (RH-2000, HIROX, Japan), based on observed structural characteristics of the ticks.

2.2 Nucleic acid extraction

Three ticks, morphologically identified as *Dermacentor* at the *genus* level, were placed in a tube and labeled according to the collecting date and lifestyles for subsequent operations. Each tube was injected with 1 mL of Stroke-Physiological Saline Solution (SPSS, 0.9% NaCl), and 3–4 0.2 mm steel balls were added before the samples were ground at a frequency of 65 Hz for 500 s. Following this, the samples were incubated in a dry water bath for 10 min at a consistent temperature of 55°C, and the supernatant fluid was carefully transferred to a new tube after a rapid centrifugation at 12,000 rpm for 1 min in a standard centrifuge. Nucleic acid extraction was then performed using the Ex-DNA/RNA nucleic acid extraction kit (TIAN LONG, China) according to the manufacturer's instructions (Shen et al., 2020), and the extracted samples were stored at −20°C for further analysis.

2.3 PCR amplification and identification

Tick identification was achieved by amplifying DNA using PCR, specifically targeting the mitochondrial gene cytochrome c oxidase I (COI; Lv et al., 2014). The primers used for the molecular biological identification of ticks were COI (Forward primer: 5′-GGAACAATATATTTAATTTTTGG-3′, Reverse primer: 5′-ATCTATCCCTACTGTAAATATATG-3′).

The PCR reactions contained 40 ng of genomic DNA and were conducted in 50 μL reaction volumes including 25 μL 2× DreamTaq PCR Master Mix (Thermo, America), 2 μL of each primer (10 μmol/L), 2 μL of DNA (20 ng/μL), and 19 μL of ddH_2O. Reactions were performed in a thermocycler under the following conditions: an initial preheating step at 95°C for 3 min, followed by 35 cycles of denaturation at 95°C for 30 s; annealing at 45°C for 30 s; extension at 72°C for 1 min; and a final extension at 72°C for 10 min.

The PCR products were subjected to 1.2% agarose gel electrophoresis (AGE) under specific conditions, including a voltage of 220 V, a current of 400 mA, and a duration of 30 min. Following this, the gel was imaged using a gel imager to visually represent the DNA fragments. For precise identification of the tick species, the nucleotide sequence of the PCR products was determined using Sanger sequencing, which was facilitated by AuGCT DNA-SYN Biotechnology Co., Ltd. located in Beijing, China.

2.4 Library preparation and sequencing

Aimed to amplify the sequences of species-specific for the 16S rRNA gene of the ribosome's large subunit, the methods followed involved the generation of a DNA library using QIAseq FX DNA Library Kit (QIAGEN, Germany), as per the manufacturer's instructions, before diluting the nucleic acid for subsequent sequencing, so that the concentration could be detected by the Qubit-4 (Thermo, America).

The primary protocol encompassed several steps, including library preparation via the fusion method, amplicon purification, quantification, and pooling. The final products were individually validated using AGE and any samples with concentrations falling below the standard 50 ng/μL, as defined by the Nextera XT Index Kit v2 (Illumina), were deemed unqualified and hence excluded from the pool. All pooled DNA samples were then subjected to paired-end sequencing, utilizing the MiSeq Reagent Kit V3 on the Illumina HiSeq platform (Odendaal et al., 2022). This sequencing process adhered strictly to the manufacturer's instructions and was characterized by an insert size of 350 bp and a read length of 250 bp.

2.5 Sequence data analysis

The barcode-based data were analyzed using the Quantitative Insights Into Microbial Ecology version2 (QIIME2) software suite (Caporaso et al., 2010), which provides a comprehensive software environment, data standards, and tool wrappers. The sequence data, presented in FASTQ format, were demultiplexed and consolidated into the same directory before input as QIIME artifacts (.qza) or QIIME visualizations (.qzv). After importing the paired-end reads from the original DNA fragments into the virtual machine, we utilized the DADA2 method to denoise data by truncating both forward and reverse sequences at 230 base pairs each. Following the DADA2 protocol, the sequences were sorted and grouped into Autosave (ASV), which is replaced by Operational Taxonomic Unit (OTU) in the following. The sequences were matched against the Silva database according to the Uchime algorithm (Edgar et al., 2011). Upon creation or acquisition of the classifier artifact, the "*qiime taxa barplot*" command was used to perform taxonomic analysis of the OTU representative sequences. The bacterial composition of each sample was visualized in a bar figure using the R packages ggplot2 and RColorBrewer (Wickham, 2009; Neuwirth, 2022).

2.6 The alpha and beta diversity analysis

For single-sample diversity analysis or alpha diversity, we determined sufficient sequencing depth based on rarefaction curves for the observed number of OTUs across all samples shown as illustrated in the Good's coverage. After the sequences of all samples were randomly drawn to a uniform data volume, we calculated the number of unique OTUs, the community richness as determined by the Shannon estimator, and the community evenness as indicated by the Pielou index in each sample using the R package Vegan (Dixon, 2003). The Wilcoxon rank-sum test was employed to assess whether these indices differed significantly and the analysis of similarities (ANOSIM) was used to quantify the variation in bacterial composition resulting from different tick lifestyles using package Vegan.

Following the exclusion of *Rickettsiaceae*, the abundance of bacteria in samples was recalculated and beta diversity analyses were performed at the *family* level using the R packages Vegan, phyloseq, and ggplot2 (McMurdie and Holmes, 2013). Weighted and unweighted UniFrac analyses were utilized to examine diversity and to compare different groups, which were subsequently plotted in a principal coordinate analysis (PCoA). The Welch's *t*-test was applied to assess differences in bacterial composition between two tick lifestyles. Variations in bacterial composition between groups were visualized using STAMP software (Parks et al., 2014). Lastly, to identify microbial taxa with significant differences between Groups A and B from *phylum* to *species* level, we employed Linear Discriminant Analysis (LDA) Effect Size (LEfSe; Segata et al., 2011). A threshold LDA score of >2 was used within 95% confidence intervals.[1]

2.7 Phylogenetic analysis and network relationship

The representative COI gene sequences and pathogen reads obtained were cross-referenced with existing results in GenBank, employing the Basic Local Alignment Search Tool (BLAST) search engine provided by the National Center for Biotechnology Information (NCBI).[2] Multiple sequence alignment was conducted using the ClustalW algorithm within the MEGA-11 software suite, applying default parameters. The phylogenetic tree was then constructed using the Kimura two-parameter model of Neighbor-Joining method based on MEGA-11, with bootstrap values estimated from 1,000 replicates.

The Spearman rank correlation test was conducted to present the interrelationship between two bacteria in samples using the R packages Hmisc and Igraph (Csardi and Nepusz, 2005), and the networks were constructed to visualize relationships between OTUs at the *genus* level. The two bacteria were linked by red or blue lines which represented a positive or negative correlativity, respectively, a process facilitated by Gephi software.[3] OTUs that were either unassigned or occurred only once were disregarded at the *genus* level and a cut-off value was established for reads to eliminate pathogens. Microbial occurrences that represented less than 10% of the frequency distribution of reads in a microorganism were discarded, and the correlation coefficient ($|r|$-value) was set at >0.6 with a 95% confidence level ($p < 0.05$).

3 Results

3.1 Tick identification

The ticks were categorized into two groups based on their mode of capture: Group A consisted of feeding ticks, while Group B comprised questing ticks. A total of 150 adult ticks were identified as *Dermacentor nuttalli* (Acari: *Ixodidae*) according to statistics, confirmed through morphological identifications and a PCR assay of

1 http://huttenhower.sph.harvard.edu/galaxy
2 http://blast.ncbi.nlm.nih.gov/blast.cgi
3 https://gephi.org

the species-specific COI region, which a high identity range of 92.66–99.75%. The phylogeny is shown in Figure 1.

3.2 Pathogen detection

Prior to the 16S metagenomics sequencing, our laboratory conducted nested PCR assays using specific primers, through which five pathogens, namely *Rickettsia*, *Anaplasma*, *Borrelia burgdorferi*, *Babesia*, and *Bartonella*, were detected. The primer sequences are outlined in Table 1, displayed from 5′ to 3′ for both forward and reverse sequences.

Of all the tested nucleic acid samples, the prevalence rates for *Rickettsia* and *Anaplasma* stood at 32.67 and 4.67%, respectively, surpassing those of the other three pathogens. Notably, *Borrelia burgdorferi*, *Babesia*, and *Bartonella* were not detected using conventional PCR assays. A subset of the samples (4.67%) demonstrated detected with *Rickettsia* and *Anaplasma* simultaneously. Results of the positive rate are presented in Supplementary Table 1.

3.3 Data general statistic

Using the Illumina HiSeq platform and after eliminating low-quality sequences, 5,480,349 reads were obtained in the data analysis. The bacterial genera were determined by comparing the taxonomic profiles at the *genus* level against the Silva database. Approximately 8×10^6 paired-end V3-V4 16S reads were procured from all samples, with the read count varying from 7,332 to 417,497 and Supplementary Table 2 included 16S NGS sample productions. After pre-processed (merged, quality filtered, and removal of singletons and chimeras) and post-processed, a cumulative total of 5 million sequences were obtained for the samples and their replicates. The average number of sequences was 109,607 and a range from 42,877 to 257,450 as illustrated in Figure 2A. The rarefaction curves, which level off well before 10,000 reads (Figure 2B), indicate that the number of reads is sufficient to compile a reliable list of bacterial genera. The raw data had uploaded to NCBI with the Bio-Project Accession Number PRJNA1015185.

3.4 Taxonomy and group comparison

At the *family* level, *Rickettsiaceae*, *Coxiellaceae*, and *Enterobacteriaceae* represented an average of 84.21, 9.42, and 3.84%, respectively. The top 20 taxonomics for each sample are illustrated in Figure 3A. Although *Rickettsiaceae* is predominant in most samples, the *Coxiellaceae* and the *Fancisellaceae* are significantly prevalent in samples number 25 (33,450 reads, 68.18%) and 50 (65,231 reads, 48.29%) respectively. As shown in Figure 3B, *Rickettsiaceae* comprises 2,330,700 (81.11%) and 2,241,292 (87.71%) of the reads in Groups A and B, respectively. Notably, *Coxiellaceae* and *Enterobacteriaceae* in Group A (339,077 reads, 11.80%; 160,804 reads, 5.60%) have approximately double and triple the reads in Group B (172,405 reads, 6.75%, 47,821

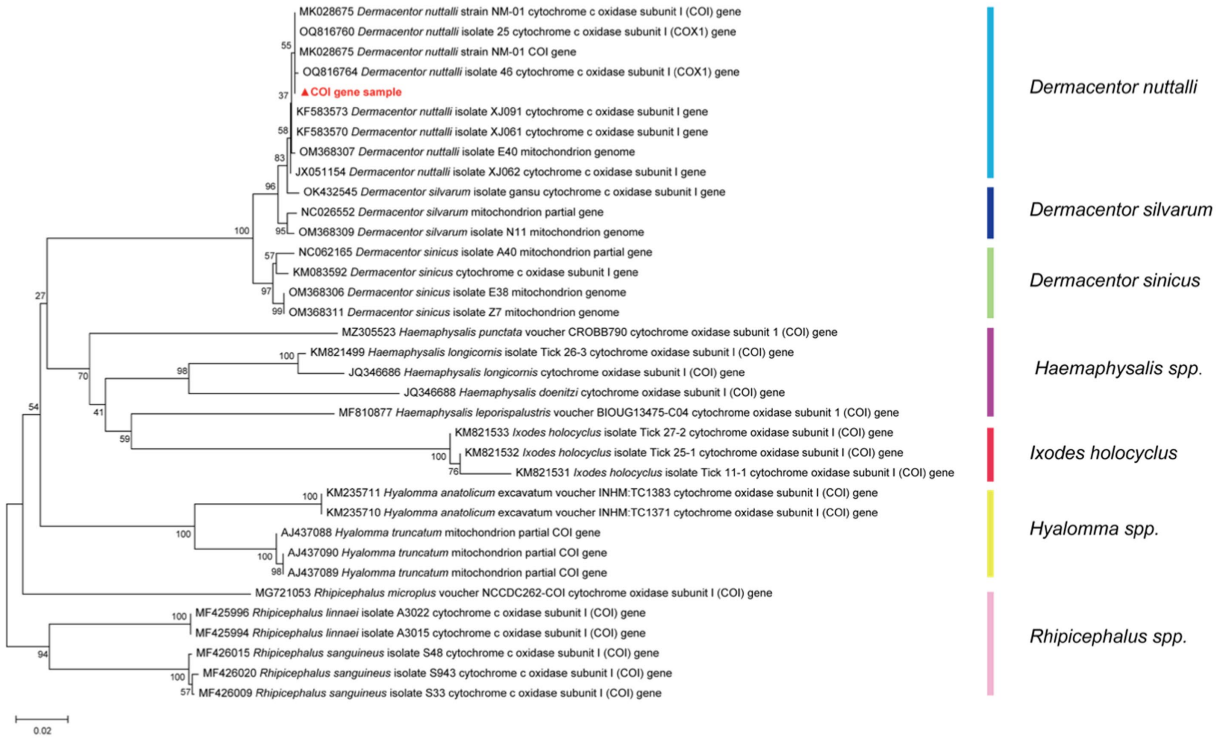


FIGURE 1
The phylogenetic tree of the COI gene with NCBI Blast results of tick species, was constructed based on the Kimura two-parameter model of Neighbor-Joining method using 1,000 repeated estimated bootstrap values.

10.3389/fmicb.2023.1264939

TABLE 1 The primer sequence of five pathogens from 5′ to 3′.

Target genes and primer sequences used for nested PCR				
Pathogen	Target gene	Primer name	Sequence	Tm (T/°C)
Rickettsia	ompA	190.70f	5′-ATGGCGAATATTTCTCCAAAA-3′	54
		190.602r	5′-AGTGCAGCATTCGCTCCCCCT-3′	
		38 s1	5′-AAAACCGCTTTATTCACC-3	
		384r1	5′-GGCAACAAGTTACCTCCT-3′	
Anaplasma	16S rRNA	Eh-out1	5′-TTGAGAGTTTGATCCTGGCTCAGAACG-3′	54
		Eh-out2	5′-CACCTCTACACTAGGAATTCCGCTATC-3″	
		Eh-gs1	5′-GTAATAACTGTATAATCCCTG-3	
		Eh-gs2	5′-GTACCGTCATTATCTTCCCTA-3″	
Borrelia burgdorferi	ITS	23 s3	5′-CGACCTTCTTCGCCTTAAAGC-3	59
		23Sa	5′-TAAGCTGACTAATACTAATTACCC-3′	
		23 s5	5′-CTGCGAGTTCGCGGGAGA-3′	
		23S6	5′-TCCTAGGCATTCACCATA-3	
Babesia	18S rRNA	Bab1	5′-CTTAGTATAAGCTTTTATACAGC-3″	54
		Bab4	5′-ATAGGTCAGAAACTTGAATGATACA-3′	
		Bab2	5′-GTTATAGTTTATTTGATGTTCGTTT-3″	
		Bab3	5′-AAGCCATGCGATTCGCTAAT-3′	
Bartonella	ITS	302F	5′-YCTTCGTTTCTCTTTCTTCA-3′	55
		473R	5′-AACCAACTGAGCTACAAGCC-3′	
		311F	5′-CTCTTTCTTCAGATGATGATCC-3′	
		448R	5′-GGATAAACCGGAAAACCTTC-3′	

reads, 1.87%) respectively. In contrast, the *Francisellaceae family*, although less prominent, shows a slight distinction between the groups, with Group B (66,723 reads, 2.61%) having more reads and proportions than Group A (1,187 reads, 0.04%).

And at the *genus* level, *Rickettsia* is the dominant *genus* across all samples, comprising between 25.36 and 98.95% of the total community (average: 82.63%), followed by *Coxiella* and *Enterobacter*, which account for averages of 9.42, and 3.49%, respectively. Figure 3C provides a visual representation of the relative abundance of the top 25 species at the *genus* level. The figure illustrates that the *genus*-level classifications of *Rickettsia* and *Ac37b* collectively constitute the *family*-level classification of *Rickettsiaceae*. The comparison of the top 25 species between groups is shown in Figure 3D.

Regarding alpha diversity, there were no significant differences between groups for either Shannon Index ($W = 359$, $p = 0.375$) or Simpson Index ($W = 353$, $p = 0.441$), despite the range of the two Index being wider for Group A compared to Group B, as presented in Supplementary Figure 1. The Wilcoxon rank test of other indexes also had no statistical differences which are shown in Supplementary Table 3. There were no significant differences in bacterial microbiome compositions between parasitic and questing ticks according to ANOSIM (pseudo-$R = 0.019$, $p = 0.175$), as illustrated previously in Figure 2F.

3.5 The diversity removed *Rickettsiaceae*

However, after removing *Rickettsiaceae* computationally, the composition unveils noteworthy variations between groups (Figure 2C). It is apparent that Figure 2D explicitly presents the dominant role of *Coxiellaceae* within two groups (A: 339,077 reads, 62.46%; B: 172,405 reads, 54.88%), and subsequently, a significant disparity in the presence of *Enterobacteriaceae* (A: 160,804 reads, 29.62%; B: 47,821 reads, 15.22%) is discernible in Figure 4. Interestingly, while the relative abundance of *Francisellaceae* in Group B (21.24%) surpasses that in Group A (0.22%), no statistical difference in microbiome composition at the *family* level is noted ($p = 0.55$, Wilcoxon rank-sum test) as depicted in Figure 2E. A comparative analysis of the *Ixodes* microbiome diversity reveals greater species evenness in Group B (0.628) relative to Group A (0.520) based on the Simpson index, while the Shannon index suggests that the species richness in Group A (1.044) is lower than in Group B (1.363) as shown in Supplementary Table 4.

Principal coordinate analysis identifies the species basically similar between the groups. The Unweighted UniFrac PCoA further reveals that PC1 (26.44%) surpasses PC2 (9.34%), and in the Weighted UniFrac PCoA plot, PC2 (36.13%) is inferior to PC1 (46.65%), as

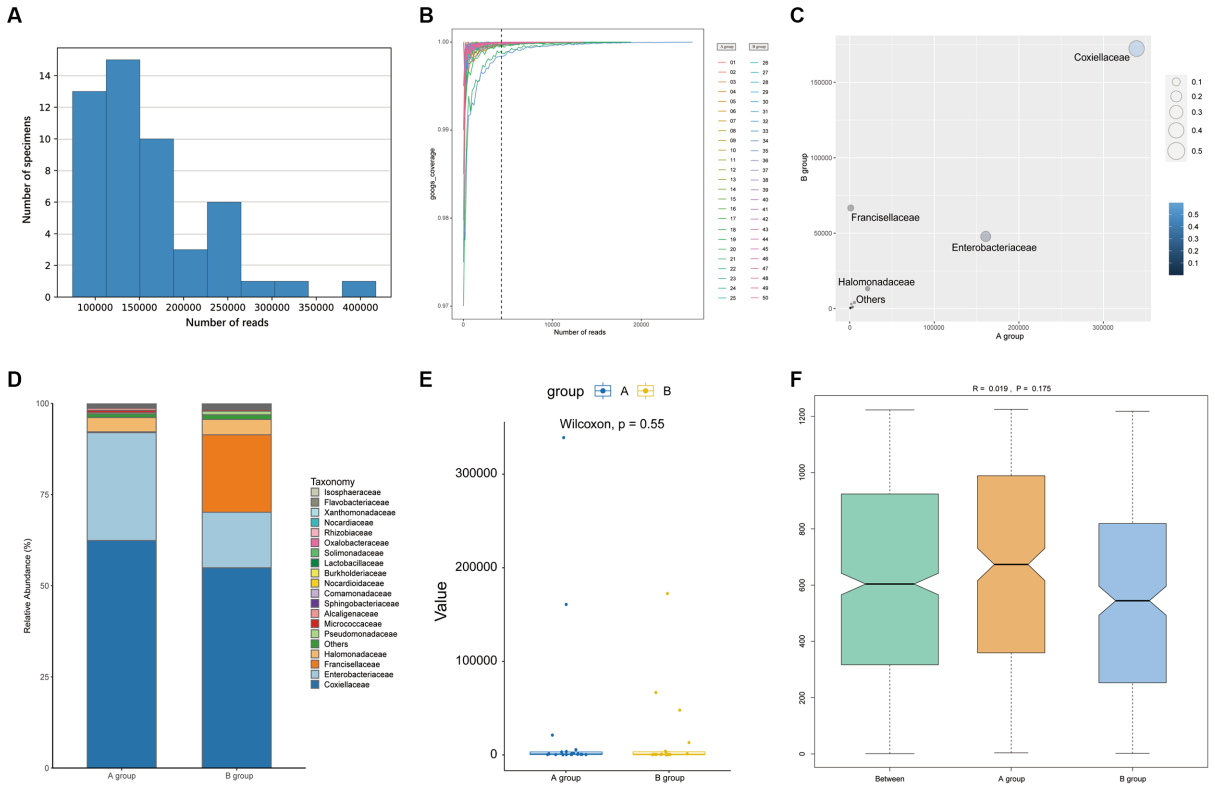


FIGURE 2
The plots show the information of reads **(A,B)** and the comparison between groups after removing *Rickettsiaceae* at the *family* level **(C–F)**. **(A)** The bar shows the reads in samples. **(B)** Rarefaction curves estimated from reads obtained. **(C)** The number of reads of each *family* of bacteria is plotted. **(D)** The bar plot shows the microbial community between groups. **(E)** A Wilcoxon test of OTUs between two groups with a box plot. **(F)** The ANOSIM analysis between groups (pseudo-R = 0.019, p = 0.175).

represented in Figures 5A,B. A comparison of the relative abundance of the top 20 bacterial genera using the Welch's *t*-test (two sides), identifies no significant differences at the *family* level within the 95% confidence intervals, as shown in Figure 6A. While *Enterobacteriaceae* is marginally more prevalent in Group A than Group B, the mean proportion of *Francisellaceae* and *Coxiellaceae* is comparatively diminished. However, substantial differences become apparent at the *genus* level as shown in Figure 6B, with the result of the relative abundance of *Friedmanniella* and *Bordetella* being elevated in parasitic ticks as compared to questing ticks. Figure 7A demonstrates differences among specimens at $p < 0.05$ as determined by LEfSe analysis, and a slight discrepancy is detectable within 95% confidence intervals from *phylum* to *species* in Figure 7B, with 11 taxa in Group B and 10 taxa in Group A detected at higher relative abundance.

3.6 Phylogenetic tree and bacterial network

For the selected representative reads from *Rickettsia, Anaplasma*, *Francisella*, and *Coxiella,* the phylogenetic tree, as shown in Figure 8, manifests sequences alignment implemented with MEGA 11 software, utilizing the Kimura two-parameter model of Neighbor-Joining method. Table 2 presents the average intraspecies pairwise genetic distances for the four *Rickettsia* and 14 reference sequences retrieved from NCBI, which range from 0.045 to 0.173 and 0.000 to 0.012, respectively. The table also indicates the pairwise divergence (Divergence = 1 − percent identity) between stains, denoted above the diagonal in Table 2.

The network, as depicted in Figure 9A, illustrates the correlational relationship among bacterial genera based on co-occurrence patterns, with varying line colors representing different relationships between two genera. The network consists of 36 nodes and 133 links, with only two instances of negative correlation noted between *Francisella* and *Phyllobacterium* ($r = -0.65$), and between *Francisella* and *Arthrobacter* ($r = -0.63$). There are 22 pairs exhibiting strong positive correlation ($r > 0.8$), with the top three pairs being *Sphingobacteriales-uncultured* and *Phyllobacterium* ($r = 0.88$), *Blastocatellia-uncultured* and *Propionibacteriaceae-uncultured* ($r = 0.92$), and *Enterobacter* and *Escherichia-Shigella* ($r = 0.98$). The detailed results for all bacteria are provided in Supplementary Table 5. In order to examine the correlation between bacterial genera across groups, we plotted network diagrams for Group A (Figure 9B) and Group B (Figure 9C). And there are 390 correlational relationships among bacteria in Group A and more than 350 in Group B, which are shown in Supplementary Tables 6, 7.

4 Discussion

Ticks, widely distributed blood-sucking parasitic arthropod, can be categorized into feeding ticks and questing ticks based on their living environment. While numerous studies have compared microbial

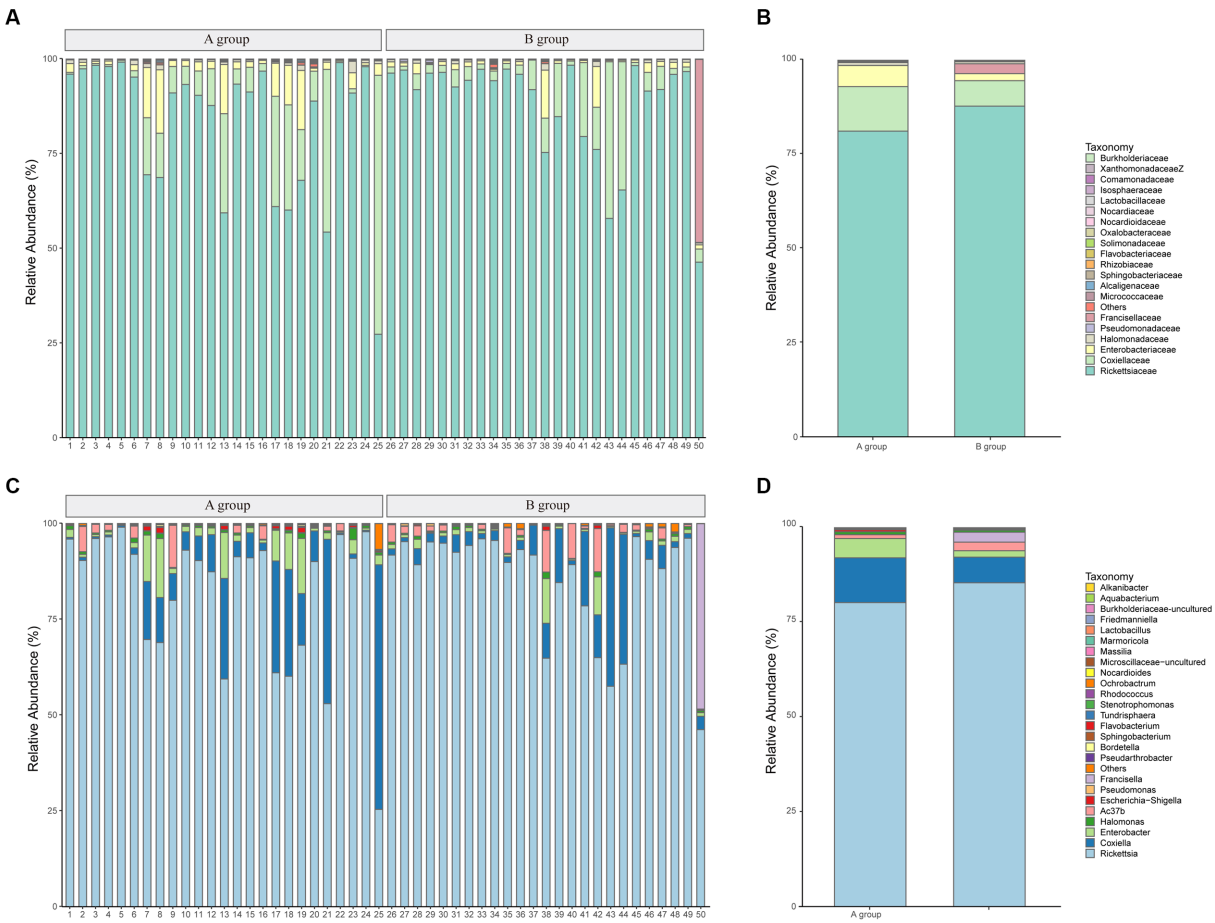


FIGURE 3
The microbial community bar plots of bacteria comparison. **(A)** The top 20 communities at the *family* level for samples. **(B)** The top 20 communities at the *family* level with groups. **(C)** The top 25 communities at the *genus* level for samples. **(D)** The top 25 communities at the *genus* level with groups.

compositions of various tick species morphologically (Krige et al., 2021), the presence of definitive differences in composition and diversity of abundance between feeding and questing ticks remains a debate topic. Our study, however, elucidates the presence of diverse compositions within groups, indicating the value in comparing two ticks of the same species and from the same sampling site.

Ixodidae species are widely distributed in Wuwei city, such as *Hyalurus asiaticus*, *Haemaphysalis qinghaiensis*, *Dermacentor nuttalli*, and *Lycoris argenteus* (Sun et al., 2016), and our molecular biological results for *Dermacentor nuttalli* align with the ecological species at the sampling site. In contrast to a study conducted in Inner Mongolia (Jiao et al., 2021), which detected the bacterial genera *Rickettsia*, *Anaplasma*, and *Coxiella* in *D. nuttalli*, our research detected the same bacteria in *D. nuttalli*, indicating that this species of *D. nuttalli* is predisposed to carry these bacteria. However, the detection rate of *Anaplasma* (4.67%, 7/150) in our results was lower than in the former study, where *Anaplasma* spp. was consistently reported and speculated to be a dominant regional species in Inner Mongolia (Jiao et al., 2021). This suggests that greater attention should be directed toward *D. nuttalli* in Wuwei City for comprehensive microorganism investigation, and the sample size, sampling points, and range of ticks should be expanded in further research.

In the relative abundance graph (Figure 3A), the *family Rikettsiaceae* constitutes the largest proportion, accounting for 81.11% in Group A and 87.71% in Group B, primarily including the genera *Rickettsia* and *Ac37b* at the *genus level*. *Rickettsia* peculiarly grows only within the cytoplasm of eucaryotic cells without producing exotoxins, and destructs host-cells from the inside, which are at high risk of infection as obligate intracellular bacteria first described by *Ricketts* in 1909 (Ricketts, 1909). After dengue infection, rickettsial diseases have been reported as the second most common cause of non-malarial febrile infectious illness in Southeast Asia (Acestor et al., 2012). One study reported a significantly higher prevalence of *Rickettsia* spp. in the *genus Dermacentor* than in the other tick species (Del Cerro et al., 2022), and a Malaysian study demonstrated that *Rickettsia RF2125* plays a dominant role in both feeding [31.8% (7/22)] and questing [25.0% (19/76)] ticks (Kho et al., 2019), which aligns with our research that *Rickettsia* exhibited the highest relative abundance and implying a propensity for *Dermacentor* to harbor *Rickettsia* spp.

The blood-feeding behavior of ticks may explain the observed statistical difference between *Friedmanniella* and *Bordetella* in detected bacteria, which may be largely attributable to the host's herbivorous traits and potential respiratory diseases. Utilizing the Welch's *t*-test and LEfSe to detect possible *genus*-level differences, we found that *Friedmanniella*, a *genus* belonging to the *Propionibacteriaceae family* frequently detected in plant samples, exhibits significant differences between groups. Group A contained

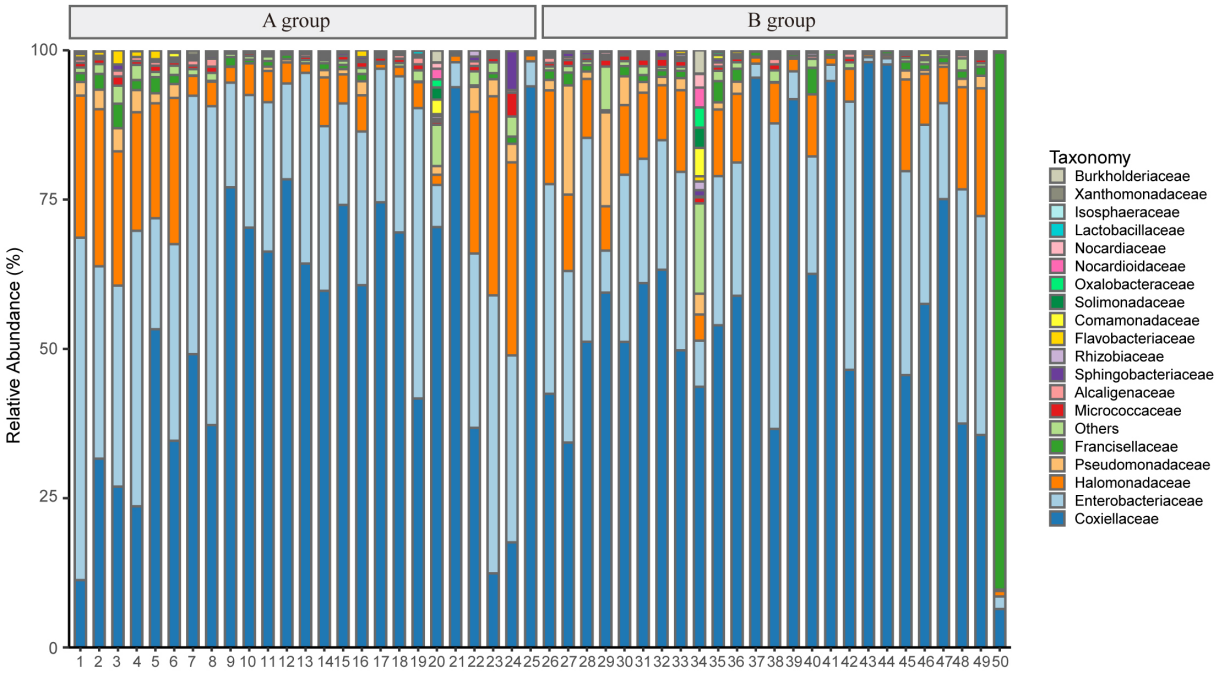


FIGURE 4
The plot shows the relative abundance of taxonomy after removing *Rickettsiaceae* at the *family* level.

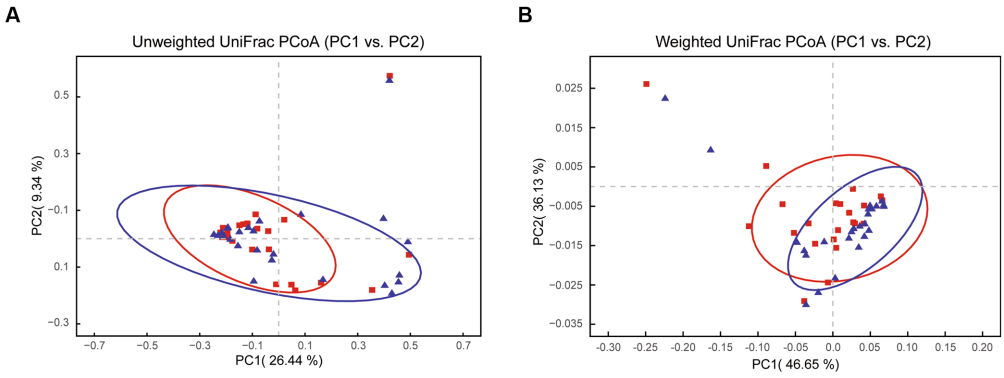


FIGURE 5
The plot shows PCoA results between groups after removing *Rickettsiaceae*. **(A)** The Unweighted UniFrac PCoA graph shows PC1 (26.44% variation) and PC2 (9.34% variation). **(B)** The Weighted UniFrac PCoA graph shows PC1 (46.65% variation) and PC2 (36.13% variation).

more reads than Group B ($p = 0.015$). Likewise, the *Bordetella*, an animal-associated *genus* known to cause whooping cough (*Bordetella pertussis*), had more reads in Group A and the difference was significant ($p = 0.030$). Extensive research into *Bordetella* has focused on the respiratory pathogens *B. pertussis* and *B. bronchiseptica* (Stevenson and Roberts, 2003), and they cause respiratory illness in children and animals including mice, dogs, pigs, and poultry, respectively (Trainor et al., 2015; Nieves and Heininger, 2016). Transmission of *Bordetellae* is typically attributed to respiratory droplets from coughing patients (Nieves and Heininger, 2016) and does not normally invade beyond the respiratory tract (Stevenson and Roberts, 2003). The bacteria can colonize in the respiratory tract, blood, and body fluids (such as cerebrospinal fluid; Liao et al., 2022), and *B. bronchiseptica* persist in a variety of environments outside the host, including water and on surfaces (Porter et al., 1991). Though no current reports manifest the presence of *Bordetella* in arthropods, our findings imply that *Dermacentor nuttalli* could be a potential vector for *Bordetella*, contradicting other reports that solely suggest transmission and infectious in mammals (Mattoo and Cherry, 2005). However, 75% ethyl alcohol may not be able to kill all pathogens on the surface of ticks, resulting in bacteria contaminating their surfaces, which requires further study to verify ticks as the vector of *Bordetella*.

Although our research detected three ticks for pathogens in one sample, it is possible that pathogens are co-infected in one tick. Tick-borne co-infections result from infection by genetically distinct pathogens (Cutler et al., 2021). In our study, the simultaneous detection rate of *Coxiella*, *Francisella*, and *Rickettsia* was 79.59% (39/49), as demonstrated in Figure 3C, and the prevalence indicates

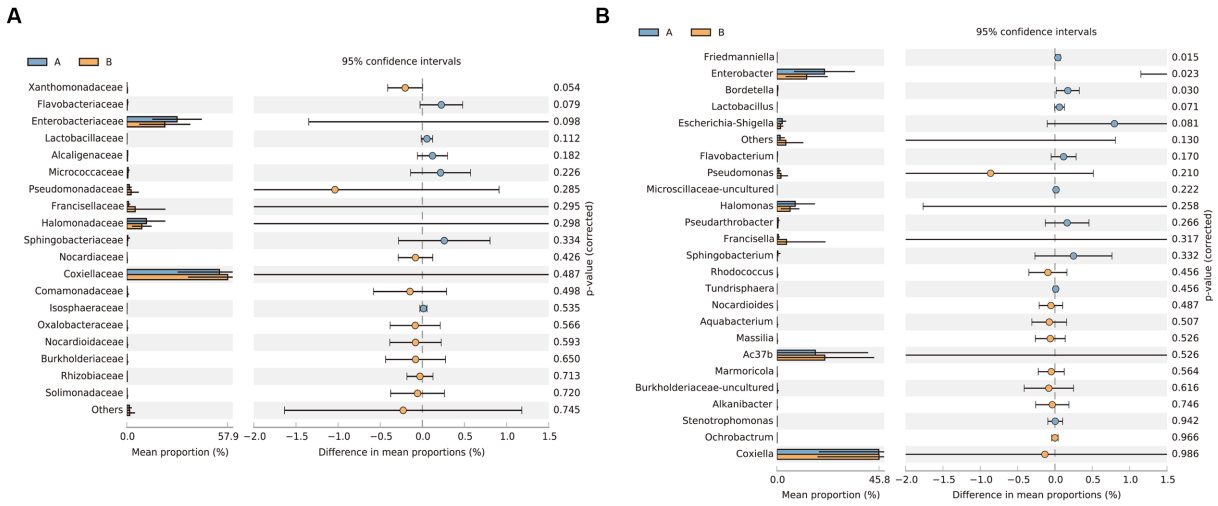


FIGURE 6
The plot shows the results of the Welch's *t*-test between groups after removing *Rickettsiaceae*. **(A)** Differences of bacterial composition in 95% confidence intervals at the *family* level. **(B)** Differences of bacterial composition in 95% confidence intervals at the *genus* level.

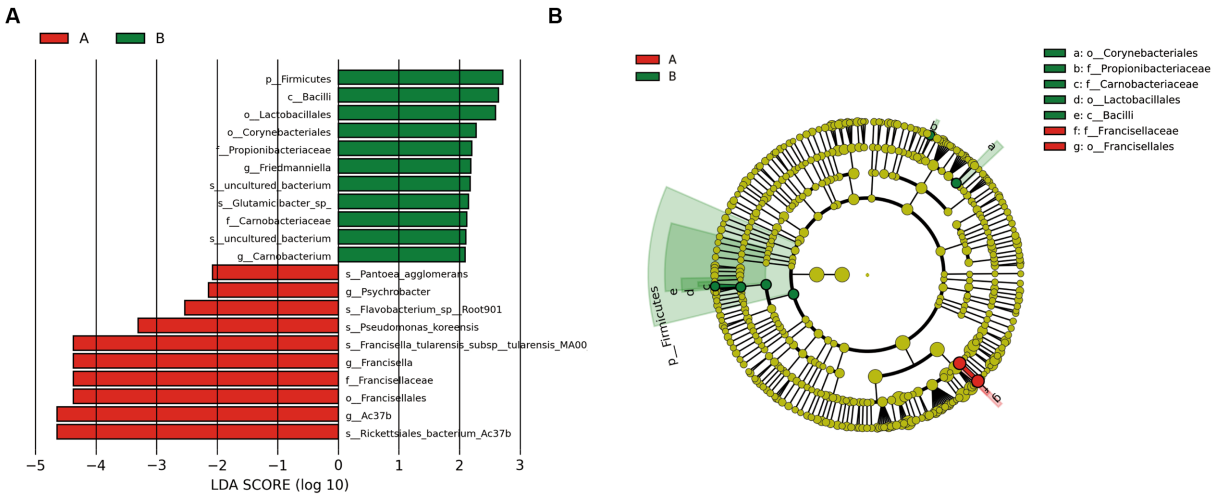


FIGURE 7
The plot shows LEFSe results between groups after removing *Rickettsiaceae*. **(A)** The bar plot explains the contribution degree of distinct species (LDA score > 2, $p < 0.05$). **(B)** The inner to outer circle of the cladistic map of distinct species evolution represents the classification level from *phylum* to *family*.

that these genera are remarkably commonplace in multiple tick species and possibly function as symbionts within ticks (Abreu et al., 2019). Indeed, co-infections and symbiote relationships are frequently observed among arthropods such as ticks, ranging from closely related variants within the same species to highly diverse pathogens, including parasites, bacteria, and viruses (Cutler et al., 2021). One research indicates a substantial presence of 90% prevalence rate (153/170) of *R. africae* in *Amblyomma variegatum* ticks and harbored a combination of pathogens, notably including *Coxiella burnetii* among others (Ehounoud et al., 2016). Furthermore, corroborative research has provided evidence of simultaneous infection with two different types of *Rickettsia* (namely *Rickettsia* spp. and *Rickettsiella* spp.) in cases (Raulf et al., 2018). Other bacteria of notable prevalence are *Coxiella* and *Francisella*, which have been reported to exist symbiotically in ticks such as *Rickettsia*-like, *Coxiella*-like, and *Francisella*-like endosymbiont (Song et al., 2022), and potentially play a crucial role in tick development by supplying B vitamins (Brenner et al., 2021).

Networks show the interrelationship between bacteria and Supplementary Tables 6, 7 present the detailed correlation results of groups. The *Francisella* appears to be inhibited or compete with 10 bacterial genera such as *Carnobacterium* ($r = -0.77$) and *Phyllobacterium* ($r = -0.70$), and the *Flavobacterium* inhibit *Friedmanniella* bacteria in Group A ($r = -0.79$). Conversely, no such competitive relations are observed in Group B. This phenomenon may be related to the competition between bacteria inherently present (such as *Francisella*) in questing ticks and those acquired from the host's blood by parasitic ticks. Notably, strong positive correlation

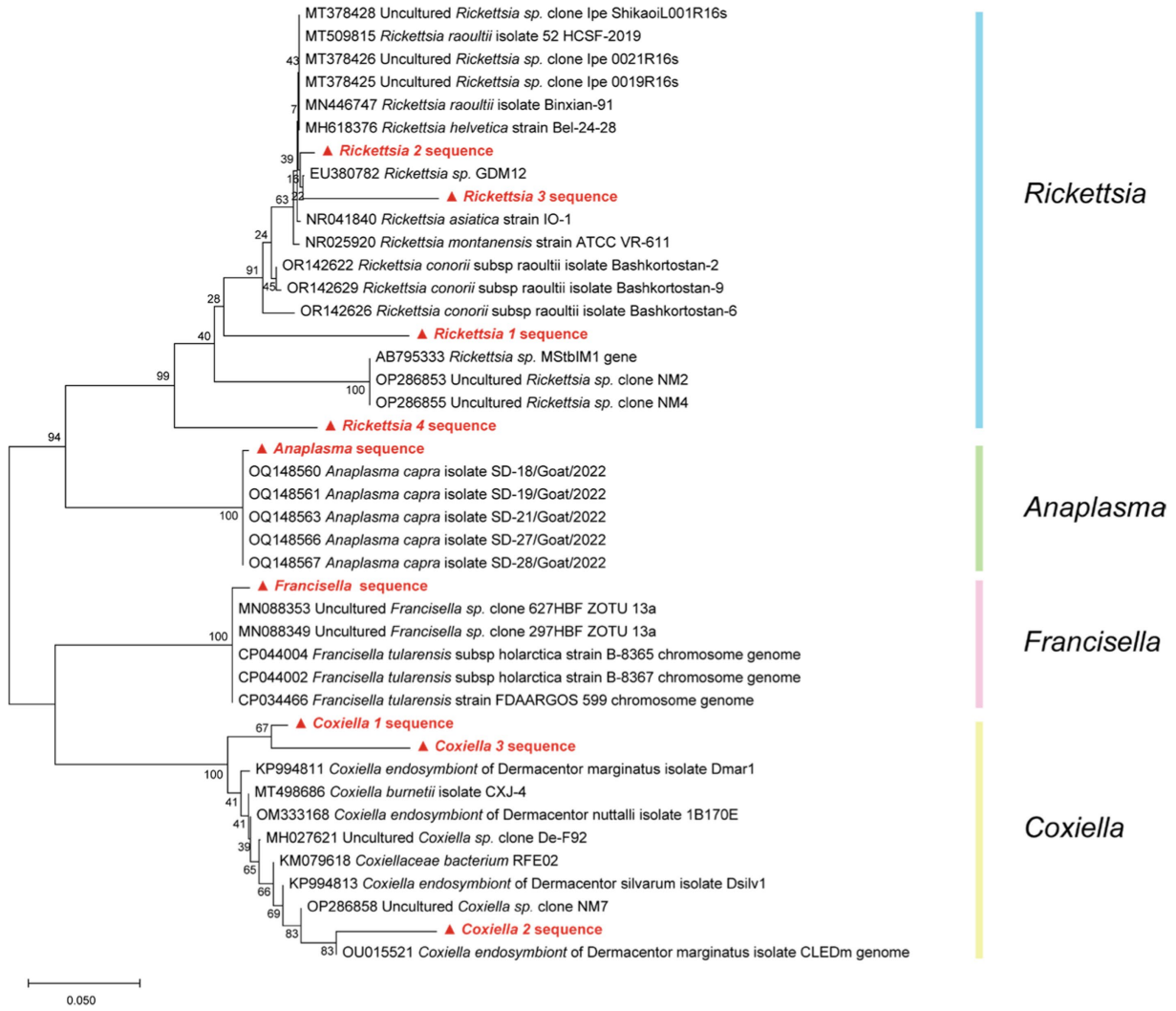


FIGURE 8
Based on the Kimura two-parameter model of Neighbor-Joining method, the phylogenetic tree of four pathogens (*Rickettsia*, *Anaplasma*, *Francisella*, and *Coxiella*) caused by NCBI blast result was drawn with 1,000 repeated estimated bootstrap values.

relationships ($r > 0.8$) are evident in both groups with the result between *Phyllobacterium* and *Arthrobacter* (Group A = 0.81, Group B = 0.86), *Phyllobacterium* and *Sphingobacteriales-uncultured* (Group A = 0.84, Group B = 0.99). And the bacteria *Halomonas*, *Escherichia-Shigella,* and *Enterobacter* have strong positive relationships in two groups, which three bacteria may promote growth mutually. It has been also reported that the three bacteria can be detected in animals simultaneously, such as cows, *Haemonchus contortus*, and *Scylla paramamosain* (Zhang et al., 2018; Chen et al., 2020; Mafuna et al., 2021). Interestingly, *Bordetella* has strong positive correlativity with the bacteria *Halomonas* ($r = 0.81$), *Enterobacter* ($r = 0.90$), and *Escherichia-Shigella* ($r = 0.92$) in Group A. Three bacteria, with more abundance in Group A, probably promoted the colonization and growth of *Bordetella*, which may related to the phenomenon that the *Bordetella* has significant differences between two groups.

Our study acknowledges that the evidence focusing on just one sampling site, one city, and the same tick species may be less sufficient. Additionally, the interactive mechanisms of symbiotic bacteria were not examined and verified. Nonetheless, our work yields significant findings. The bacterial genera *Friedmanniella* and *Bordetella* have statistical differences between parasitic and questing ticks, and the interrelationships in bacteria are different in diverse lifestyles ticks. In conclusion, our comparison of parasitic and questing ticks provides unique perspective into diversities in pathogen proportion and microbial communities, leveraging the 16S rRNA gene sequencing and analysis capabilities of NGS. In further research, the sampling range should be expanded and the mechanisms between bacterial genera warrant resolution and optimization.

5 Conclusion

The conducted experimental investigations have elucidated differences in the causative agents present in two lifestyle categories of ticks within Wuwei City. It was found that microorganisms demonstrate significant variations between parasitic and questing ticks, with the potential for bacterial genera to either inhibit or

TABLE 2 Genetic distance (above the diagonal) and pairwise divergence (below the diagonal) in the *Rickettsia* sequences with NCBI reference sequence based on the phylogenetic tree.

	Rickettsiaceae .1	*Rickettsiaceae* .2	*Rickettsiaceae* .3	*Rickettsiaceae* .4	MG 827281	MF 496154	LC 602357	MZ 292055	EU 380782	NR 041840	JQ 480832	NR 025920	LC 089861	MF 002589	MT 509813	OL 423537	MT 279305	OQ 392416
Rickettsiaceae.1	—	0.886	0.862	0.847	0.9	0.903	0.903	0.903	0.9	0.9	0.898	0.898	0.893	0.903	0.903	0.903	0.903	0.808
Rickettsiaceae.2	0.125	—	0.954	0.879	0.976	0.978	0.978	0.978	0.976	0.976	0.978	0.973	0.968	0.973	0.973	0.973	0.973	0.806
Rickettsiaceae.3	0.152	0.045	—	0.859	0.956	0.959	0.959	0.959	0.956	0.956	0.959	0.954	0.956	0.954	0.954	0.954	0.954	0.818
Rickettsiaceae.4	0.173	0.135	0.155	—	0.893	0.896	0.896	0.896	0.893	0.893	0.9	0.9	0.891	0.896	0.896	0.896	0.898	0.91
MG827281	0.108	0.025	0.043	0.116	—	0.998	0.998	0.998	0.995	0.995	0.993	0.993	0.988	0.993	0.993	0.993	0.993	0.825
MF496154	0.105	0.022	0.04	0.114	0.002	—	1	1	0.998	0.998	0.995	0.995	0.99	0.995	0.995	0.995	0.995	0.828
LC602357	0.105	0.022	0.04	0.114	0.002	0	—	1	0.998	0.998	0.995	0.995	0.99	0.995	0.995	0.995	0.995	0.828
MZ292055	0.105	0.022	0.04	0.114	0.002	0	0	—	0.998	0.998	0.995	0.995	0.99	0.995	0.995	0.995	0.995	0.828
EU380782	0.108	0.025	0.043	0.116	0.005	0.002	0.002	0.002	—	0.995	0.993	0.993	0.988	0.993	0.993	0.993	0.993	0.828
NR041840	0.108	0.025	0.043	0.116	0.005	0.002	0.002	0.002	0.005	—	0.993	0.993	0.988	0.993	0.993	0.993	0.993	0.83
JQ480832	0.111	0.022	0.04	0.108	0.007	0.005	0.005	0.005	0.007	0.007	—	0.995	0.99	0.995	0.995	0.995	0.995	0.828
NR025920	0.111	0.028	0.045	0.108	0.007	0.005	0.005	0.005	0.007	0.007	0.005	—	0.99	0.995	0.995	0.995	0.995	0.833
LC089861	0.117	0.033	0.043	0.119	0.012	0.01	0.01	0.01	0.012	0.012	0.01	0.01	—	0.99	0.99	0.99	0.99	0.833
MF002589	0.105	0.028	0.045	0.114	0.007	0.005	0.005	0.005	0.007	0.007	0.005	0.005	0.01	—	1	1	0.995	0.83
MT509813	0.105	0.028	0.045	0.114	0.007	0.005	0.005	0.005	0.007	0.007	0.005	0.005	0.01	0	—	1	0.995	0.83
OL423537	0.105	0.028	0.045	0.114	0.007	0.005	0.005	0.005	0.007	0.007	0.005	0.005	0.01	0	0	—	0.995	0.83
MT279305	0.105	0.028	0.045	0.111	0.007	0.005	0.005	0.005	0.007	0.007	0.005	0.005	0.01	0.005	0.005	0.005	—	0.83
OQ392416	0.215	0.219	0.205	0.088	0.193	0.189	0.189	0.189	0.189	0.186	0.189	0.183	0.183	0.186	0.186	0.186	0.186	—

Divergence = 1 – percent identity.

10.3389/fmicb.2023.1264939

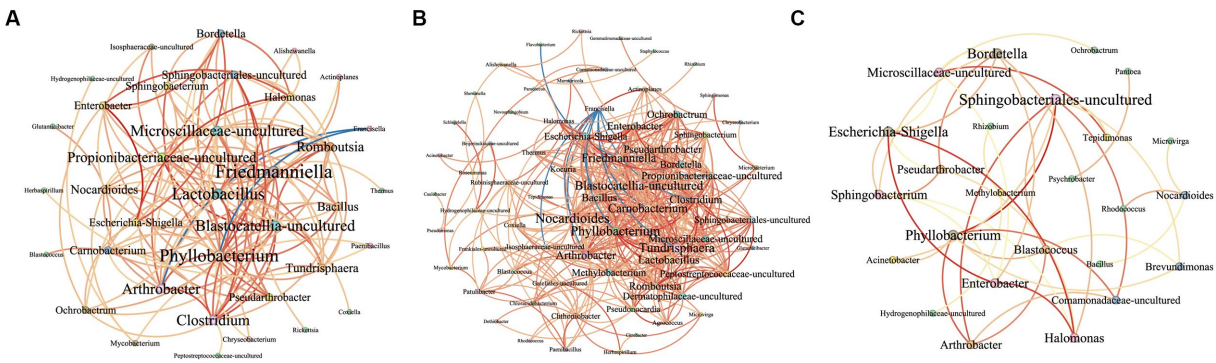


FIGURE 9
This image presents an interrelationship network among various bacterial genera at the *genus* level. Negative and positive correlations are represented by blue and red lines, respectively. The gradation from yellow to red in the lines signifies the degree of correlation, with a deeper red indicating a stronger correlation. The size of the bacterial names corresponds to the number of relationships each has within the network, with larger names indicating a greater number of connections. **(A)** The network shows the correlations of bacteria in all samples. **(B)** The network shows the correlations of bacteria in Group A. **(C)** The network shows the correlations of bacteria in Group B.

promote each other within diverse tick populations. These results underscore significance of lifestyle-based classification in researching potential pathogens.

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 below: NCBI—PRJNA1015185.

Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study.

Author contributions

LZ: Data curation, Formal analysis, Writing – original draft, Writing – review & editing. JH: Formal analysis, Investigation. QZ: Methodology, Supervision. ZH: Methodology, Writing – review & editing. S-WS: Methodology, Resources, Writing – review & editing. RL: Data curation, Resources, Writing – review & editing. R-SL: Investigation, Visualization, Writing – review & editing. W-KZ: Project administration, Validation, Writing – review & editing. Y-HW: Methodology, Software, Writing – review & editing. L-LX: Investigation, Supervision, Writing – review & editing. Z-HL: Investigation, Software, Writing – review & editing. Z-JS: 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 a grant from the National Natural Science Foundation of China (Grant No. 82273689).

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.2023.1264939/full#supplementary-material

References

Abreu, D. P. B., Peixoto, M. P., Luz, H. R., Zeringóta, V., Santolin, Í., Famadas, K. M., et al. (2019). Two for the price of one: co-infection with Rickettsia bellii and spotted fever group Rickettsia in Amblyomma (Acari: Ixodidae) ticks recovered from wild birds in Brazil. *Ticks Tick Borne Dis* 10:101266. doi: 10.1016/j.ttbdis.2019.101266

Acestor, N., Cooksey, R., Newton, P. N., Ménard, D., Guerin, P. J., Nakagawa, J., et al. (2012). Mapping the aetiology of non-malarial febrile illness in Southeast Asia through a systematic review--terra incognita impairing treatment policies. *PLoS One* 7:e44269. doi: 10.1371/journal.pone.0044269

Brenner, A. E., Muñoz-Leal, S., Sachan, M., Labruna, M. B., and Raghavan, R. (2021). Coxiella burnetii and related tick endosymbionts evolved from pathogenic ancestors. *Genome Biol. Evol.* 13. doi: 10.1093/gbe/evab108

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Chaorattanakawee, S., Wofford, R. N., Takhampunya, R., Katherine Poole-Smith, B., Boldbaatar, B., Lkhagvatseren, S., et al. (2022). Tracking tick-borne diseases in Mongolian livestock using next generation sequencing (NGS). *Ticks Tick Borne Dis* 13:101845. doi: 10.1016/j.ttbdis.2021.101845

Chen, H., Fu, K., Pang, B., Wang, J., Li, H., Jiang, Z., et al. (2020). Determination of uterine bacterial community in postpartum dairy cows with metritis based on 16S rDNA sequencing. *Vet. Anim. Sci.* 10:100102. doi: 10.1016/j.vas.2020.100102

Ciebiera, O., Łopińska, A., and Gabryś, G. (2021). Ticks on game animals in the fragmented agricultural landscape of western Poland. *Parasitol. Res.* 120, 1781–1788. doi: 10.1007/s00436-021-07132-9

Csardi, G., and Nepusz, T. (2005). The igraph software package for complex network research. *InterJ. Complex Syst.* 1695, 1–9.

Cutler, S. J., Vayssier-Taussat, M., Estrada-Peña, A., Potkonjak, A., Mihalca, A. D., and Zeller, H. (2021). Tick-borne diseases and co-infection: current considerations. *Ticks Tick Borne Dis* 12:101607. doi: 10.1016/j.ttbdis.2020.101607

Del Cerro, A., Oleaga, A., Somoano, A., Barandika, J. F., García-Pérez, A. L., and Espí, A. (2022). Molecular identification of tick-borne pathogens (Rickettsia spp., *Anaplasma phagocytophilum*, *Borrelia burgdorferi* sensu lato, Coxiella burnetii and piroplasms) in questing and feeding hard ticks from North-Western Spain. *Ticks Tick Borne Dis* 13:101961. doi: 10.1016/j.ttbdis.2022.101961

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14, 927–930. doi: 10.1111/j.1654-1103.2003.tb02228.x

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Ehounoud, C. B., Yao, K. P., Dahmani, M., Achi, Y. L., Amanzougaghene, N., Kacou N'Douba, A., et al. (2016). Multiple pathogens including potential new species in tick vectors in Côte d'Ivoire. *PLoS Negl. Trop. Dis.* 10:e0004367. doi: 10.1371/journal.pntd.0004367

Jiao, J., Lu, Z., Yu, Y., Ou, Y., Fu, M., Zhao, Y., et al. (2021). Identification of tick-borne pathogens by metagenomic next-generation sequencing in Dermacentor nuttalli and *Ixodes persulcatus* in Inner Mongolia, China. *Parasit. Vectors* 14:287. doi: 10.1186/s13071-021-04740-3

Kahl, O. (2018). Hard ticks as vectors-some basic issues. *Wien. Klin. Wochenschr.* 130, 479–483. doi: 10.1007/s00508-018-1360-x

Kho, K. L., Tan, P. E., and Tay, S. T. (2019). Diversity of Rickettsiae in feeding and questing ticks collected from a Malaysian Forest reserve area. *J. Med. Entomol.* 56, 547–552. doi: 10.1093/jme/tjy168

Krige, A. S., Thompson, R. C. A., Seidlitz, A., Keatley, S., Wayne, J., and Clode, P. L. (2021). Molecular detection of Trypanosoma spp. in questing and feeding ticks (Ixodidae) collected from an endemic region of south-West Australia. *Pathogens* 10:1037. doi: 10.3390/pathogens10081037

Lan-Hua, L., and Yi, Z. (2019). Prevalence and control of tick-borne parasitic diseases in China. *Chin. J. Schistosom. Control* 31, 58–62. doi: 10.16250/j.32.1374.2018310

Liao, Y., Li, W. R., Zhu, Y., Luo, S. H., Liao, Q., and Wan, C. M. (2022). Invasive *Bordetella pertussis* infection in infants: a case report. *Open Forum Infect. Dis.* 9:ofac478. doi: 10.1093/ofid/ofac478

Lv, J., Wu, S., Zhang, Y., Chen, Y., Feng, C., Yuan, X., et al. (2014). Assessment of four DNA fragments (COI, 16S rDNA, ITS2, 12S rDNA) for species identification of the Ixodida (Acari: Ixodida). *Parasit. Vectors* 7:93. doi: 10.1186/1756-3305-7-93

Mafuna, T., Soma, P., Tsotetsi-Khambule, A. M., Hefer, C. A., Muchadeyi, F. C., Thekisoe, O. M. M., et al. (2021). Bacterial profiling of Haemonchus contortus gut microbiome infecting Dohne merino sheep in South Africa. *Sci. Rep.* 11:5905. doi: 10.1038/s41598-021-85282-7

Mattoo, S., and Cherry, J. D. (2005). Molecular pathogenesis, epidemiology, and clinical manifestations of respiratory infections due to *Bordetella pertussis* and other *Bordetella* subspecies. *Clin. Microbiol. Rev.* 18, 326–382. doi: 10.1128/CMR.18.2.326-382.2005

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes. R package version 1.1-3. Available at: https://CRAN.R-project.org/package=RColorBrewer

Nieves, D. J., and Heininger, U. (2016). Bordetella pertussis. *Microbiol Spectr* 4:21. doi: 10.1128/microbiolspec.EI10-0008-2015

Odendaal, M. L., Groot, J. A., Hasrat, R., Chu, M., Franz, E., Bogaert, D., et al. (2022). Higher off-target amplicon detection rate in MiSeq v3 compared to v2 reagent kits in the context of 16S-rRNA-sequencing. *Sci. Rep.* 12:16489. doi: 10.1038/s41598-022-20573-1

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Porter, J. F., Parton, R., and Wardlaw, A. C. (1991). Growth and survival of *Bordetella bronchiseptica* in natural waters and in buffered saline without added nutrients. *Appl. Environ. Microbiol.* 57, 1202–1206. doi: 10.1128/aem.57.4.1202-1206.1991

Raulf, M. K., Jordan, D., Fingerle, V., and Strube, C. (2018). Association of Borrelia and Rickettsia spp. and bacterial loads in *Ixodes ricinus* ticks. *Ticks Tick Borne Dis* 9, 18–24. doi: 10.1016/j.ttbdis.2017.10.014

Ricketts, H. T. (1909). A MICRO-organism which apparently has a specific relationship to ROCKY mountain spotted fever: a PRELIMINARY REPORT. *J. Am. Med. Assoc.* LII, 379–380. doi: 10.1001/jama.1909.25420310039002,M.D

Rochlin, I., and Toledo, A. (2020). Emerging tick-borne pathogens of public health importance: a mini-review. *J. Med. Microbiol.* 69, 781–791. doi: 10.1099/jmm.0.001206

Rodino, K. G., Theel, E. S., and Pritt, B. S. (2020). Tick-borne diseases in the United States. *Clin. Chem.* 66, 537–548. doi: 10.1093/clinchem/hvaa040

Rodino, K. G., Wolf, M. J., Sheldon, S., Kingry, L. C., Petersen, J. M., Patel, R., et al. (2021). Detection of tick-borne Bacteria from whole blood using 16S ribosomal RNA gene PCR followed by next-generation sequencing. *J. Clin. Microbiol.* 59. doi: 10.1128/JCM.03129-20

Sanschagrin, S., and Yergeau, E. (2014). Next-generation sequencing of 16S ribosomal RNA gene amplicons. *J. Vis. Exp.* 90. doi: 10.3791/51709

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Shen, L., Huang, F., Chen, X., Xiong, Z., Yang, X., Li, H., et al. (2020). Diagnostic efficacy of three test kits for SARS-CoV-2 nucleic acid detection. *Zhejiang Da Xue Xue Bao Yi Xue Ban* 49, 185–190. doi: 10.3785/j.issn.1008-9292.2020.03.09

Song, X., Zhong, Z., Gao, L., Weiss, B. L., and Wang, J. (2022). Metabolic interactions between disease-transmitting vectors and their microbiota. *Trends Parasitol.* 38, 697–708. doi: 10.1016/j.pt.2022.05.002

Stevenson, A., and Roberts, M. (2003). Use of Bordetella bronchiseptica and *Bordetella pertussis* as live vaccines and vectors for heterologous antigens. *FEMS Immunol. Med. Microbiol.* 37, 121–128. doi: 10.1016/S0928-8244(03)00068-3

Sun, M., Guan, G., Liu, Z., Wang, J., Wang, D., Wang, S., et al. (2020). Molecular survey and genetic diversity of *Babesia* spp. and *Theileria* spp. in cattle in Gansu Province, China. *Acta Parasitol.* 65, 422–429. doi: 10.2478/s11686-020-00179-1

Sun, M., Wang, S. F., Wang, D. Q., Yang, Y. B., and Ma, C. (2016). Preliminary investigation on species and distribution of ticks and tick-borne diseases in Wuwei City, Gansu Province. *Anim. Quarant. China* 33, 23–25. doi: 10.3969/j.issn.1005-944X.2016.02.010

Sun, Y., Xu, R., Liu, Z., Wu, M., and Qin, T. (2019). Ornithodoros (Ornithodoros) huajianensis sp. nov. (Acari, argasidae), a new tick species from the Mongolian marmot (*Marmota bobak* sibirica), Gansu province in China. *Int J Parasitol Parasites Wildlife* 9, 209–217. doi: 10.1016/j.ijppaw.2019.05.001

Trainor, E. A., Nicholson, T. L., and Merkel, T. J. (2015). *Bordetella pertussis* transmission. *Pathog Dis* 73:ftv068. doi: 10.1093/femspd/ftv068

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer

Zhang, X., Sun, Z., Zhang, X., Zhang, M., and Li, S. (2018). Hemolymph microbiomes of three aquatic invertebrates as revealed by a new cell extraction method. *Appl. Environ. Microbiol.* 84. doi: 10.1128/AEM.02824-17

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