frontiers | Research Topics

# Evolutionary mechanisms of infectious diseases, volume II

**Edited by**
Jianying Gu, Yufeng Wang and Zhan Zhou

**Published in**
Frontiers in Microbiology

ISSN 1664-8714
ISBN 978-2-8325-2192-2
DOI 10.3389/978-2-8325-2192-2

## About Frontiers

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

## Frontiers journal series

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

## Dedication to quality

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

## What are Frontiers Research Topics?

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

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

# Evolutionary mechanisms of infectious diseases, volume II

**Topic editors**

Jianying Gu — College of Staten Island, United States
Yufeng Wang — University of Texas at San Antonio, United States
Zhan Zhou — Zhejiang University, China

**Citation**

Gu, J., Wang, Y., Zhou, Z., eds. (2023). *Evolutionary mechanisms of infectious diseases, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2192-2

# Table of contents

05 Editorial: Evolutionary mechanisms of infectious diseases, volume II
Zhan Zhou, Jianying Gu and Yufeng Wang

08 Characterization of a Type VI Secretion System *vgrG2* Gene in the Pathogenicity of *Burkholderia thailandensis* BPM
Jin Li, Wei-wei Hu, Guo-xin Qu, Xiao-rong Li, Yi Xiang, Peng Jiang, Jiang-qiao Luo, Wen-huan He, Yu-jia Jin and Qiong Shi

18 Global Phylogeny of *Mycobacterium avium* and Identification of Mutation Hotspots During Niche Adaptation
Rachel Mizzi, Karren M. Plain, Richard Whittington and Verlaine J. Timms

34 Genomic Insights Into the Interspecific Diversity and Evolution of *Mobiluncus*, a Pathogen Associated With Bacterial Vaginosis
Yisong Li, Ying Wang and Jie Liu

46 Identification of region of difference and H37Rv-related deletion in *Mycobacterium tuberculosis complex* by structural variant detection and genome assembly
Zhuochong Liu, Zhonghua Jiang, Wei Wu, Xinyi Xu, Yudong Ma, Xiaomei Guo, Senlin Zhang and Qun Sun

60 Complete genome sequences of classical swine fever virus: Phylogenetic and evolutionary analyses
Yue Liu, Amina Nawal Bahoussi, Pei-Hua Wang, Changxin Wu and Li Xing

68 CircDDX17 enhances coxsackievirus B3 replication through regulating miR-1248/NOTCH receptor 2 axis
Tingjun Liu, Yuhan Li, Shengjie Chen, Lulu Wang, Xiaolan Liu, Qingru Yang, Yan Wang, Xiaorong Qiao, Jing Tong, Xintao Deng, Shihe Shao, Hua Wang and Hongxing Shen

80 Surveillance of avian influenza viruses in live bird markets of Shandong province from 2013 to 2019
Ti Liu, Yousong Peng, Julong Wu, Shangwen Lu, Yujie He, Xiyan Li, Lin Sun, Shaoxia Song, Shengyang Zhang, Zhong Li, Xianjun Wang, Shu Zhang, Mi Liu and Zengqiang Kou

90 Comparative transcriptomic analyzes of human lung epithelial cells infected with wild-type SARS-CoV-2 and its variant with a 12-bp missing in the E gene
Yi-Sheng Sun, Hao Sun, Han-Ping Zhu, Gao-Lei Li, Fang Xu, Hang-Jing Lu, An Tang, Bei-Bei Wu, Yu-Dong Li, Ping-Ping Yao and Jian-Min Jiang

100 Mutation-driven parallel evolution in emergence of ACE2-utilizing sarbecoviruses
Bin Gao and Shunyi Zhu

116 **Molecular evolution, diversity, and adaptation of foot-and-mouth disease virus serotype O in Asia**
Fangtao Li, Yan Li, Jianrong Ma, Ruizhi Wu, Xingqi Zou, Yebing Liu, Qizu Zhao and Yuanyuan Zhu

122 **Variation in synonymous evolutionary rates in the SARS-CoV-2 genome**
Qianru Sun, Jinfeng Zeng, Kang Tang, Haoyu Long, Chi Zhang, Jie Zhang, Jing Tang, Yuting Xin, Jialu Zheng, Litao Sun, Siyang Liu and Xiangjun Du

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 03 April 2023
DOI 10.3389/fmicb.2023.1192566

# Editorial: Evolutionary mechanisms of infectious diseases, volume II

Zhan Zhou[1]*, Jianying Gu[2]* and Yufeng Wang[3]*

[1]Institute of Drug Metabolism and Pharmaceutical Analysis, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China, [2]Department of Biology, College of Staten Island, City University of New York, Staten Island, New York, NY, United States, [3]Department of Molecular Microbiology Immunology, South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX, United States

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Ludmila Chistoserdova,
University of Washington, United States

*CORRESPONDENCE
Zhan Zhou
✉ zhanzhou@zju.edu.cn
Jianying Gu
✉ Jianying.gu@csi.cuny.edu
Yufeng Wang
✉ yufeng.wang@utsa.edu

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 23 March 2023
ACCEPTED 24 March 2023
PUBLISHED 03 April 2023

CITATION
Zhou Z, Gu J and Wang Y (2023) Editorial: Evolutionary mechanisms of infectious diseases, volume II.
*Front. Microbiol.* 14:1192566.
doi: 10.3389/fmicb.2023.1192566

KEYWORDS

**evolution, infectious diseases, infection, pathogenicity, virulence**

Editorial on the Research Topic
Evolutionary mechanisms of infectious diseases, volume II

Infectious diseases continue to pose significant global health challenges, as they are among the leading causes of mortality worldwide, according to the World Health Organization (WHO) (WHO, 2020). The ongoing emergence and re-emergence of infectious pathogens highlight the complexity and dynamism of these diseases, underscoring the need for proactive and adaptive approaches to their management and control (Morens et al., 2004; Morse et al., 2012; Morens and Fauci, 2020; Frutos et al., 2021; Baker et al., 2022).

Infectious diseases are dynamic and constantly evolving, driven by a variety of evolutionary mechanisms that enable pathogens to adapt, survive, and spread within host populations. Understanding these mechanisms is crucial for developing effective prevention and control strategies. Recent advances in the study of "evolutionary mechanisms of infectious diseases" have seen significant progress (Woolhouse et al., 2005; Sironi et al., 2015; Geoghegan and Holmes, 2018; Gomez-Carballa et al., 2020; Seitz et al., 2020; Cao et al., 2022; Obermeyer et al., 2022). Scientists have unraveled the complex interactions between pathogen evolution, antibiotic resistance, and host adaptation through in-depth analysis of pathogen genomes, with the help of high throughput omics technologies and big data analytical approaches (Didelot et al., 2016; Grote and Earl, 2022; Zhou et al., 2023). Pathogens rapidly adapt to environmental changes *via* mechanisms such as mutation, genetic recombination, and horizontal gene transfer (Frost et al., 2005; Shi et al., 2022). Furthermore, research on host immune systems and microbiomes helps understand disease transmission and outbreak dynamics (Virgin, 2014; Zheng et al., 2020). These developments have spurred the creation of novel vaccines and antibiotics to combat the growing threat of infectious diseases (Rappuoli and Aderem, 2011; Excler et al., 2021). In the future, interdisciplinary research and collaboration will enhance prediction, prevention, and control of disease transmission (Morse et al., 2012; Zeng et al., 2021).

The second volume of our Research Topic, "*Evolutionary mechanisms of infectious diseases*, volume II," builds on the foundation established by the first volume (Gu et al., 2021), further delving into the complex interplay between pathogens, hosts, and the environment. This collection of articles expands our understanding of the evolutionary processes underlying infectious diseases and highlights the importance of a multidisciplinary approach to tackle the challenges that they present. By studying the evolutionary mechanisms, researchers can gain valuable insights into the processes driving the emergence, spread, and persistence of infectious diseases (Woolhouse et al., 2005; Geoghegan and Holmes, 2018). This knowledge can inform the development of more effective prevention and control strategies, such as targeted vaccination campaigns (Andre et al., 2008; Kuehn, 2022), antimicrobial stewardship programs (Dyar et al., 2017; Rice, 2018), and surveillance systems to monitor and respond to emerging and re-emerging pathogens (Morse et al., 2012; Baker et al., 2022).

We sincerely thank all contributors of our Research Topic. This collection of 11 articles is divided into four sections. The first section includes four articles presenting comprehensive genomic analyses of viral and bacterial pathogens, such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), avian influenza viruses, classical swine fever viruses, and Mobiluncus, offering insights into their evolution, transmission dynamics, and interactions with host immune systems. (1) Sun Q. et al. explore the impact of synonymous mutations on the SARS-CoV-2 genome, revealing their potential functional correlations and providing valuable insights for better pandemic control by analyzing the synonymous evolutionary rate and its relationship with host proteins. (2) Liu, T. et al. present long-term surveillance of avian influenza viruses (AIVs) in the live bird market in Shandong province, identifying H9 as the predominant subtype and providing insights into the epidemic and evolution of AIVs, which can inform more effective control strategies in China. (3) Liu Y. et al. reanalyze 203 complete genomic sequences of classical swine fever viruses (CSFVs), identifying new lineages and potential natural recombination events involving vaccine and highly virulent strains, emphasizing the need for careful vaccine applications and alternative preventive strategies for better CSFV management. (4) Li Y. et al. present the first genome-level analysis of *Mobiluncus*, a pathogen linked to bacterial vaginosis, uncovering phylogenetic distinctions, functional diversification, and evolutionary dynamics that could contribute to better understanding and treatment of the infection.

The second section presents in-depth investigations into the role of non-coding RNAs, such as circular RNAs and long non-coding RNAs, in the regulation of viral replication, host immunity, and the development of novel therapeutic targets. Sun Y-S. et al. compare the transcriptome profiles of mRNA and lncRNAs in human lung epithelial cells infected with different SARS-CoV-2 strains, identifying differentially expressed genes and pathways that may explain the varying replication and immunogenicity properties of the strains, thus offering insights into the pathogenesis of SARS-CoV-2 variants. Liu, T. et al. reveal that circDDX17, a circular RNA, promotes Coxsackievirus B3 (CVB3) replication and regulates NOTCH2 by acting as a miRNA sponge for miR-1248, offering new insights into the role of non-coding RNAs in viral infections.

The third section is composed by four articles focused on the impact of mutations, recombination events, and other genetic factors on pathogen evolution, emphasizing the importance of understanding these mechanisms to inform the development of effective prevention and control strategies. (1) Gao and Zhu investigate the origin of ACE2 binding in sarbecoviruses, a group of evolutionarily related β-coronaviruses including SARS-CoV-2, suggesting that three distinct ancestral RBDs independently developed the ACE2 binding trait through parallel mutations, providing insights into the mutation-driven evolution of sarbecoviruses in their early history. (2) Li F. et al. reveal the molecular evolution, diversity, and host tropisms of Foot-and-Mouth Disease Virus (FMDV) Serotype O in Asia, finding that the Cathay topotype has evolved at a higher rate and displayed increased genetic diversity, becoming a more severe epidemic in recent years, with a distinct host preference compared to other topotypes. (3) Mizzi et al. investigate genetic diversity of *Mycobacterium avium* subspecies, finding that unique coding sequences and mutation hotspots may serve as biomarkers for understanding virulence mechanisms and host/tissue specificity, which could lead to new diagnostic targets and advances in epidemiology and therapeutics. (4) Liu Z. et al. identify novel lineage-specific large sequence polymorphisms in *Mycobacterium tuberculosis* complex, providing insights into the genealogical differentiation and aiding in the development of stable genetic markers for genotyping.

The last section of the collection includes investigations into the complex interplay between pathogens and host immune responses, highlighting the need for a comprehensive understanding of host-pathogen interactions in the context of infectious disease evolution and control. Li J. et al. reveal that the vgrG2 gene in *Burkholderia thailandensis* plays a critical role in its pathogenicity, interaction with host cells, and host immune response, providing new insights into the bacterium's virulence mechanisms.

The articles featured in "*Evolutionary mechanisms of infectious diseases, volume II*" showcase the power of a multidisciplinary approach in deepening our understanding of infectious disease dynamics. As we continue to confront the challenges posed by emerging and re-emerging pathogens, fostering collaboration and innovation across disciplines is more critical than ever.

We extend our heartfelt gratitude to all the authors who contributed their valuable research to this topic, and to the reviewers for their diligent and insightful evaluations. We hope that this collection of articles will serve as a valuable resource for researchers, inspire further scientific inquiry, and contribute to the global effort to combat infectious diseases.

## Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Acknowledgments

We sincerely thank all researchers who have contributed to our Research Topic.

## 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

Andre, F. E., Booy, R., Bock, H. L., Clemens, J., Datta, S. K., John, T. J., et al. (2008). Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bull. World Health Organ.* 86, 140–146. doi: 10.2471/BLT.07.040089

Baker, R. E., Mahmud, A. S., Miller, I. F., Rajeev, M., Rasambainarivo, F., Rice, B. L., et al. (2022). Infectious disease in an era of global change. *Nat. Rev. Microbiol.* 20, 193–205. doi: 10.1038/s41579-021-00639-z

Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., et al. (2022). BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* 608, 593–602. doi: 10.1038/s41586-022-04980-y

Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W., and Wilson, D. J. (2016). Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* 14, 150–162. doi: 10.1038/nrmicro.2015.13

Dyar, O. J., Huttner, B., Schouten, J., and Pulcini, C. (2017). What is antimicrobial stewardship? *Clin. Microbiol. Infect.* 23, 793–798. doi: 10.1016/j.cmi.2017.08.026

Excler, J. L., Saville, M., Berkley, S., and Kim, J. H. (2021). Vaccine development for emerging infectious diseases. *Nat. Med.* 27, 591–600. doi: 10.1038/s41591-021-01301-0

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732. doi: 10.1038/nrmicro1235

Frutos, R., Gavotte, L., Serra-Cobo, J., Chen, T., and Devaux, C. (2021). COVID-19 and emerging infectious diseases: the society is still unprepared for the next pandemic. *Environ. Res.* 202, 111676. doi: 10.1016/j.envres.2021.111676

Geoghegan, J. L., and Holmes, E. C. (2018). The phylogenomics of evolving virus virulence. *Nat. Rev. Genet.* 19, 756–769. doi: 10.1038/s41576-018-0055-5

Gomez-Carballa, A., Bello, X., Pardo-Seco, J., Martinon-Torres, F., and Salas, A. (2020). Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res.* 30, 1434–1448. doi: 10.1101/gr.266221.120

Grote, A., and Earl, A. M. (2022). Within-host evolution of bacterial pathogens during persistent infection of humans. *Curr. Opin. Microbiol.* 70, 102197. doi: 10.1016/j.mib.2022.102197

Gu, J., Zhou, Z., and Wang, Y. (2021). Editorial: evolutionary mechanisms of infectious diseases. *Front. Microbiol.* 12, 667561. doi: 10.3389/fmicb.2021.667561

Kuehn, B. M. (2022). Targeted flu vaccination campaigns needed for certain racial and ethnic groups. *JAMA* 328, 2005. doi: 10.1001/jama.2022.18487

Morens, D. M., and Fauci, A. S. (2020). Emerging pandemic diseases: how we got to COVID-19. *Cell* 182, 1077–1092. doi: 10.1016/j.cell.2020.08.021

Morens, D. M., Folkers, G. K., and Fauci, A. S. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature* 430, 242–249. doi: 10.1038/nature02759

Morse, S. S., Mazet, J. A., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., et al. (2012). Prediction and prevention of the next pandemic zoonosis. *Lancet* 380, 1956–1965. doi: 10.1016/S0140-6736(12)61684-5

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., et al. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* 376, 1327–1332. doi: 10.1126/science.abm1208

Rappuoli, R., and Aderem, A. (2011). A 2020 vision for vaccines against HIV, tuberculosis and malaria. *Nature* 473, 463–469. doi: 10.1038/nature10124

Rice, L. B. (2018). Antimicrobial stewardship and antimicrobial resistance. *Med. Clin. North Am.* 102, 805–818. doi: 10.1016/j.mcna.2018.04.004

Seitz, B. M., Aktipis, A., Buss, D. M., Alcock, J., Bloom, P., Gelfand, M., et al. (2020). The pandemic exposes human nature: 10 evolutionary insights. *Proc. Natl. Acad. Sci. U. S. A.* 117, 27767–27776. doi: 10.1073/pnas.2009787117

Shi, A., Fan, F., and Broach, J. R. (2022). Microbial adaptive evolution. *J. Ind. Microbiol. Biotechnol.* 49, kuab076. doi: 10.1093/jimb/kuab076

Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* 16, 224–236. doi: 10.1038/nrg3905

Virgin, H. W. (2014). The virome in mammalian physiology and disease. *Cell* 157, 142–150. doi: 10.1016/j.cell.2014.02.032

WHO. (2020). *The Top 10 Causes of Death.* Available online at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed March 20, 2023).

Woolhouse, M.E. J., Haydon, D. T., and Antia, R. (2005). Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol.* 20, 238–244. doi: 10.1016/j.tree.2005.02.009

Zeng, D., Cao, Z., and Neill, D. B. (2021). "Chapter 22 - artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control," in *Artificial Intelligence in Medicine*, eds L. Xing, M. L. Giger, and J. K. Min (Netherlands: Academic Press; Elsevier Inc), 437–453. doi: 10.1016/B978-0-12-821259-2.00022-3

Zheng, D., Liwinski, T., and Elinav, E. (2020). Interaction between microbiota and immunity in health and disease. *Cell Res.* 30, 492–506. doi: 10.1038/s41422-020-0332-7

Zhou, B., Zhou, H., Zhang, X., Xu, X., Chai, Y., Zheng, Z., et al. (2023). TEMPO: a transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Comput. Biol. Med.* 152, 106264. doi: 10.1016/j.compbiomed.2022.106264

ORIGINAL RESEARCH
published: 05 January 2022
doi: 10.3389/fmicb.2021.811343

# Characterization of a Type VI Secretion System *vgrG2* Gene in the Pathogenicity of *Burkholderia thailandensis* BPM

*Jin Li[1], Wei-wei Hu[2], Guo-xin Qu[3], Xiao-rong Li[1], Yi Xiang[1], Peng Jiang[1], Jiang-qiao Luo[1], Wen-huan He[1], Yu-jia Jin[1] and Qiong Shi[1]**

*[1] M.O.E. Key Laboratory of Laboratory Medicine Diagnostics, Department of Laboratory Medicine, Chongqing Medical University, Chongqing, China, [2] Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [3] Department of Orthopedic Surgery, The First Affiliated Hospital of Hainan Medical University, Hainan, China*

*Burkholderia thailandensis* is a clinically underestimated conditional pathogen in the genus *Burkholderia*, the pathogenicity of the infection caused by *B. thailandensis* remains poorly understood. According to previous studies, Type-VI secretion system (T6SS) is a protein secreting device widely existing in Gram-negative bacilli. Valine-glycine repeat protein G (VgrG) is not only an important component of T6SS, but also a virulence factor of many Gram-negative bacilli. In one of our previous studies, a unique T6SS *vgrG* gene (*vgrG2* gene) was present in a virulent *B. thailandensis* strain BPM (BPM), but not in the relatively avirulent *B. thailandensis* strain E264 (E264). Meanwhile, transcriptome analysis of BPM and E264 showed that the *vgrG2* gene was strongly expressed in BPM, but not in E264. Therefore, we identified the function of the *vgrG2* gene by constructing the mutant and complemented strains in this study. *In vitro*, the *vgrG2* gene was observed to be involved in the interactions with host cells. The animal model experiment showed that the deletion of *vgrG2* gene significantly led to the decrease in the lethality of BPM and impaired its ability to trigger host immune response. In conclusion, our study provides a new perspective for studying the pathogenicity of *B. thailandensis* and lays the foundation for discovering the potential T6SS effectors.

**Keywords: BPM, pathogenicity, virulence factor, T6SS, *vgrG2* gene**

OPEN ACCESS

***Edited by:***
*Yufeng Wang,*
*University of Texas at San Antonio, United States*

***Reviewed by:***
*Qingpi Yan,*
*Jimei University, China*
*Kumutha Malar Vellasamy,*
*University of Malaya, Malaysia*

****Correspondence:***
*Qiong Shi*
*shiqiong@cqmu.edu.cn*

***Specialty section:***
*This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology*

***Received:*** *08 November 2021*
***Accepted:*** *13 December 2021*
***Published:*** *05 January 2022*

***Citation:***
*Li J, Hu W-w, Qu G-x, Li X-r, Xiang Y, Jiang P, Luo J-q, He W-h, Jin Y-j and Shi Q (2022) Characterization of a Type VI Secretion System vgrG2 Gene in the Pathogenicity of Burkholderia thailandensis BPM. Front. Microbiol. 12:811343. doi: 10.3389/fmicb.2021.811343*

## INTRODUCTION

*Burkholderia thailandensis* is a clinically underestimated conditional pathogen in the genus *Burkholderia*. It is very similar to *Burkholderia pseudomallei* in terms of colony morphology, immunogenicity and antimicrobial susceptibility (Brett et al., 1998; Ngamdee et al., 2015; Garcia, 2017). To date, only a few studies have described the isolation of *B. thailandensis* from invasive

**Abbreviations:** BPM, *Burkholderia thailandensis* BPM; *B. pseudomallei*, *Burkholderia pseudomallei*; *E. coli*, *Escherichia coli*; *P. aeruginosa*, *Pseudomonas aeruginosa*.

human infections (Lertpatanasuwan et al., 1999; Glass et al., 2006; Ginther et al., 2015; Gee et al., 2018), and the pathogenic mechanism of the infections caused by *B. thailandensis* remains poorly understood. Therefore, it is necessary to study the pathogenesis of *B. thailandensis* to more effectively prevent infections caused by *B. thailandensis*. In one of our previous studies, the virulent *B. thailandensis* strain BPM was isolated from a blood and sputum specimen of a 76-year-old man with sepsis who died in China (Chang et al., 2017). The biological and biochemical characteristics of BPM are very similar to those of *B. pseudomallei*, and the clinical symptoms and imaging findings of patients infected with this strain are consistent with acute melioidosis, manifesting as acute high fever, productive cough with white sputum and breathing difficulty (Wiersinga et al., 2012; Tang et al., 2018; Gassiep et al., 2020). However, biochemical identification test results showed that BPM was positive for arabinose assimilation, which is consistent with the biochemical characteristics of *B. thailandensis*. To confirm the biochemical identification results, 16S rRNA gene sequence and whole-genome sequencing analyses were performed. The results showed that the sequence of the 16S rRNA gene was 100% consistent with that of the *B. thailandensis* E264 (E264) 16S rRNA gene (GenBank No. CP008785.1, CP008786.1) and that the sequence obtained by whole-genome sequencing was more than 96% homologous to the genome sequence of E264 (GenBank No. CP000085.1, CP000086.1). Finally, the BPM strain was identified as *B. thailandensis* based on the NT nucleic acid sequence database, and the genome sequences of BPM have been deposited in the GenBank database under accession numbers CP050020-CP050021. In one of our previous studies, we compared the virulence of BPM and E264 in BALB/c mice, and the results indicated that the virulence of BPM was significantly higher than that of E264, which confirmed that BPM is a virulent pathogen (Chang et al., 2017). Additionally, an integrated type VI secretion system (T6SS) gene cluster was found in BPM by bioinformatic analysis. However, the pathogenicity of the T6SS involved in BPM infection is poorly understood.

The T6SS is an important virulence factor that plays a key role in microbial competition and bacterial infection (Fridman et al., 2020). It can deliver toxic effectors to bacterial and eukaryotic neighbors and plays an important role in both bacterial competition and virulence (Hsieh et al., 2019). Valine-glycine repeat protein G (VgrG) has been reported to be an important component of the functional T6SS of *B. pseudomallei* and is involved in the occurrence of acute melioidosis (Schwarz et al., 2014). However, the functions of the T6SS *vgrG* gene in the development of *B. thailandensis* infections remain unknown. In one of our previous studies, a unique T6SS *vgrG* gene (*vgrG2* gene) was present in a virulent BPM, but not in the relatively avirulent E264. Meanwhile, transcriptome analysis of BPM and E264 showed that the *vgrG2* gene was strongly expressed in BPM, but not in E264. However, the function of *vgrG2* gene remains unknown. Therefore, this study investigated the function of the putative T6SS *vgrG2* gene in BPM by knocking out the *vgrG2* gene and producing a *vgrG2* gene complementation construct. The mutant and complemented strains will be used to determine the function of the T6SS *vgrG2* gene by investigating the changes in the virulence of BPM. Altogether, this study aimed to lay a foundation for discovering potential T6SS effectors of *B. thailandensis* and provide a new perspective on the study of host cell signal transduction and immune defense mechanisms.

# MATERIALS AND METHODS

## Bacterial Strains and Growth Conditions

All bacterial strains used in this study are listed in **Table 1**. All strains were cultured on sheep blood agar plates (Thermo, United States) at 37°C for 16–20 h, and isolated colonies were inoculated into 10 mL Luria-Bertani (LB) broth (Solarbio, China), which was then stirred at 37°C for 12 h. The working cultures were prepared by transferring 100 μL of a 12 h culture to 10 mL LB broth (1:100 dilution), which was then allowed to stand at 37°C for 8 h. The stationary-phase bacteria were diluted to $10^6 \sim 10^8$ colony-forming units (CFU)/mL in LB broth, and phenotypic characteristics were evaluated (Jiang et al., 2016). These final suspensions were plated onto LB agar to accurately determine the number of CFUs per milliliter.

## Transcriptomic Analysis

The total RNA of BPM and E264 was extracted with TRIzol reagent (Invitrogen, United States). The quantity and purity of the extracted RNA were assessed using a NanoDrop ND-1000 spectrophotometer (Thermo, United States). RNA-seq libraries were created using the Illumina TruSeq Stranded mRNA Library Prep Kit (Illumina, Inc., United States) according to the manufacturer's protocol. Sequencing was performed at Shenzhen Hai-yi Biotechnology Co., Ltd. using an Illumina MiSeq System benchtop sequencing instrument (read length: 75 bp, read type: paired end) (Zhu et al., 2016). The raw sequence data were filtered by removing reads containing adapters, reads containing poly N sequences, and low-quality reads. The clean reads were aligned to the genomes of BPM and E264 by using Bowtie2-2.2.3 (Kovacs-Simon et al., 2019). DESeq was used to identify differentially expressed genes (Zhu et al., 2016).

**TABLE 1** | Bacterial strains used in this study.

| Strain | Description | Source or reference |
|---|---|---|
| BPM | A hypervirulent strain isolated from a deceased patient with *B. thailandensis* infection, $TC^S$, $Cm^S$ | Laboratory collection |
| BE264 | An environmental strain from the American Type Culture Collection, $TC^S$, $Cm^S$ | United States (ATCC 700388) |
| Δ*vgrG2* | Mutant with BPMhun02934 gene deleted in BPM, $TC^S$, $Cm^S$ | This study |
| Δ*vgrG2*/p*vgrG2* | Mutant Δ*vgrG2* complemented with gene *vgrG2*, $TC^S$, $Cm^S$ | This study |

## Construction of Mutant and Complemented Strains

To test the role of the T6SS *vgrG2* gene in the pathogenesis of *B. thailandensis* and its contribution to the development of *B. thailandensis* infection, knockout mutants of a key component (*vgrG2*) of the T6SS were constructed by double crossover recombination through allelic replacement of the suicide plasmid pLP12cm as described previously (Luo et al., 2015). The knockout mutant was designated Δ*vgrG2*. The *vgrG2* gene was amplified from the *B. thailandensis* BPM genome and then ligated into plasmid pTac-tetM to construct the complementation expression plasmid pTac-tetM-*vgrG2*. Finally, the complementation plasmids were transferred into mutants to generate complemented strains (Δ*vgrG2*/p*vgrG2*). All mutant and complemented strains were verified using PCR (**Supplementary Figure 1**) and DNA sequencing (data not shown).

## Growth Characteristics and Antimicrobial Susceptibility Testing

The strain was cultured on sheep blood agar at 37°C for 18–20 h and then transferred to LB broth for shaking culture at 180 rpm at 37°C. The growth characteristics of the BPM, mutant, and complemented strains were determined via optical density measurements (Eppendorf BioPhotometer, Germany) performed at 600 nm (OD600), and colony formation units (CFUs) were counted over a 24-h period as described previously (Zhu et al., 2020). Then, the antimicrobial susceptibilities of the BPM, mutant, and complemented strains were initially tested with a Vitek-2 Compact automatic microbiological assay system (BioMérieux, French). The experimental methods were performed according to the guidelines of the Clinical and Laboratory Standards Institute (CLSI) for *P. aeruginosa* (Bobenchik et al., 2017). Fresh bacterial colonies extracted directly from sheep blood agar were incubated at 37°C for 18–24 h and then resuspended in sterile saline to obtain a suspension of 0.5 McFarland turbidity. *E. coli* ATCC 25922 and *P. aeruginosa* ATCC 2785 were used as quality controls. The antimicrobial susceptibility testing results were explained in accordance with the CLSI M45 guidelines for *B. pseudomallei*. Each assay was performed three times.

## Animal Model Experiments

All animal experiments were approved by the research board of the Ethics Committee of the Third Military Medical University under permit number AMUMEC-20201085. To determine the 50% lethal dose ($LD_{50}$), five-week-old, pathogen-free, female BALB/c mice were obtained from Daping Hospital Animal Center. Ten BALB/c mice were used as a sample population for the survival rate of BALB/c mice infected with BPM, mutant, and complemented strains. Phosphate-buffered saline (PBS) was used as negative control. Ten BALB/c mice were selected for each bacterial concentration to determine the $LD_{50}$. Two-fold serial dilution of the bacteria was performed from a starting concentration of $8 \times 10^7$ CFU/mL to $5 \times 10^6$ CFU/mL, and BALB/c mice were infected intravenously with 0.1 mL of each concentration. Symptoms and mortality rates were observed for seven days. The exact inoculation dose was confirmed on LB agar, and the $LD_{50}$ was calculated as described by Barnes (Barnes and Ketheesan, 2005).

## Histopathological Studies

To examine the differences in the pathological changes caused by the tested strains, livers and lungs were collected from BALB/C mice infected with the BPM, mutant and complemented strains at designated times (4, 8, 12, and 16 h post infection). Tissue samples were fixed in 10% buffered formalin. Paraffin-embedded tissue sections were stained with hematoxylin and eosin according to the standard protocol and examined by light microscopy (Zhao et al., 2011).

## Systemic Measurement of Inflammatory Cytokines

To assess the function of the *vgrG2* gene in inflammation, serum samples (infected with BPM, mutant, and complemented strains) were collected, and the levels of IL-1β, IL-6 and TNF-α were measured using Mouse Precoated ELISA kits (Dakewei Biotech Co., Ltd). Each assay was performed three times.

## Whole-Blood Bactericidal Experiments

Human whole-blood samples used in the experiment were taken from 10 healthy individuals. The whole blood bactericidal assay was performed as previously described with minor modifications (Zong et al., 2019). Briefly, a bacterial inoculum of 100 μL (adjusted to $10^6$ CFU/mL) prepared from the mid-log phase was diluted with PBS and added to 900 μL of fresh whole blood contained in 24-well plates (Corning, United States), and the mixtures were incubated at 37°C. After incubation for 3 h, the bacteria were plated onto LB agar and counted. The survival rates of the BPM, mutant, and complemented strains were expressed by using the following formula: $(\mathrm{CFU/mL})_{\mathrm{t=3h}}/(\mathrm{CFU/mL})_{\mathrm{t=0h}} \times 100\%$. Each assay was performed three times.

## Cell Invasion and Survival Assays

The cell invasion assay was similar to that previously performed (Pijuan et al., 2019). RAW264.7 cells were incubated at 37°C with 5% $CO_2$ in 24-well plates at a concentration of $5 \times 10^5$ cells per well. RAW264.7 cells were grown on DMEM (Gibco GlutaMAX$^{TM}$, United States) containing glucose, glutamine, and 10% fetal bovine serum. *B. thailandensis* suspensions were added to the cells at an MOI of 10 or 100, followed by centrifugation at 500 g for 5 min and incubation at 37°C with 5% $CO_2$ for 1 h to determine invasion. One-hour post infection (hpi), the monolayers were washed twice with PBS and lysed with 0.1% Triton X-100 (Sigma, United States) in PBS, and serial dilutions were plated and incubated at 37°C for 48 h. The invasion percentages of the BPM, mutant, and complemented strains were calculated as follows: (invasion CFU/total inoculum CFU) × 100 (Lewis et al., 2017). To determine intracellular survival after initial invasion, after 1 h, the monolayers were washed twice with PBS and replenished with complete medium

**TABLE 2 |** Expression of *vgrG* gene in BPM.

| GeneID | E264_expr | BPM_expr | $\log_2$ Fold | p value | q value | Diff |
|---|---|---|---|---|---|---|
| BPM01336 | 100.6032 | 48.7003 | −1.267132379 | 1.43E-52 | 3.02E-51 | Down |
| BPM02934 | 0 | 44.2911 | 11.34104488 | 9.68E-22 | 6.33E-21 | Up |
| BPM03563 | 2.0684 | 1.492 | −0.709531306 | 0.079886227 | 0.115328157 | - |
| BPM03564 | 2.9132 | 2.1304 | −0.691301852 | 0.042559877 | 0.064857862 | - |
| BPM03921 | 17.5007 | 13.6585 | −0.580339578 | 0.000169016 | 0.000357747 | - |
| BPM04575 | 0.7314 | 1.0029 | 0.186528779 | 0.709326414 | 0.761237862 | - |
| BPM05231 | 4.877 | 6.4538 | 0.204416147 | 0.386627009 | 0.460374412 | - |
| BPM05382 | 1.0218 | 1.5411 | 0.344760554 | 0.39838535 | 0.472371201 | - |
| BPM05892 | 91.3012 | 45.9427 | −1.221455825 | 3.57E−32 | 3.70E−31 | Down |

*The first column is the gene ID; the second and third column are the standardized expressions of E264 and BPM; the fourth column is the ratio of normalized expression (BPM/E264, $\log_2$ fold change in transcriptome); the fifth column is the corrected p-value; the sixth column is the corrected q-value; the seventh column indicates genetic differences, Up indicates up-regulation, Down indicates down-regulation, noDEG indicates no difference.*

containing 250 μg/ml chloromycetin. At 3 h post inoculation, the monolayers were washed twice with PBS and then lysed with 0.1% Triton X-100 in PBS, and serial dilutions were plated and incubated at 37°C for 48 h. The percent survival of the BPM, mutant, and complemented strains was calculated as (survival CFU/invasion CFU) × 100 (Lewis et al., 2017). Each assay was performed three times.

## Cell Counting Kit 8 Assays

The cytotoxicity of the bacteria to RAW264.7 cells was tested by CCK-8 assays. Bacteria in the stationary phase resuspended in fresh medium were added to 96-well plates (Corning, United States) (MOI = 10). RAW264.7 cells were washed with PBS, resuspended in DMEM and plated in 96-well plates at a concentration of 5000 cells/well. Next, CCK-8 assay kit (MCE, China) reagents were added to the wells according to the manufacturer's instructions. The optical density at 450 nm was measured using a microplate reader (Thermo Fisher Scientific, Varioskan LUX, China) to assess cell viability. Cytotoxicity was expressed according to the following formula: cytotoxicity (%) = (test sample − low control)/(high control − low control) × 100 (Tang et al., 2021). Each assay was performed three times.

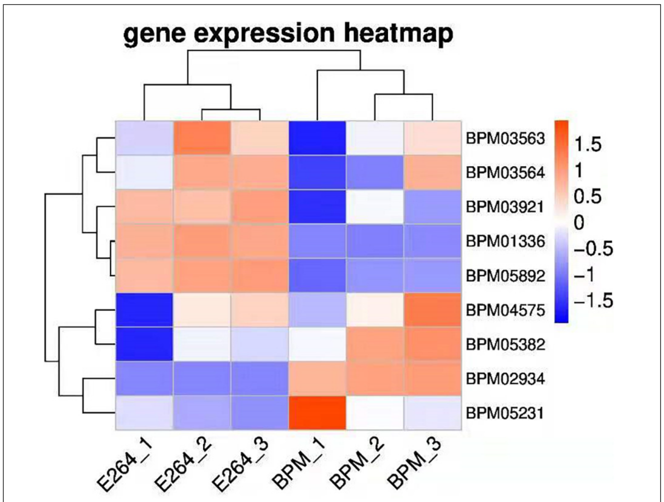


**FIGURE 1 |** Heatmap of *vgrG* gene expression in BPM and E264. Colors in the heatmap represent gene expression levels among samples. The BPM and E264 results came from three repeated samples.

## Statistics

Statistical analyses were performed using GraphPad Prism 7 (San Diego, United States). One-way ANOVA with the log-rank test was used to compare BPM to the mutant and complemented strains. We also used Tukey's multiple comparison test to compare each strain to all other strains. Significant differences between groups are indicated: * ($P < 0.05$), ** ($P < 0.01$) and *** ($P < 0.001$).

# RESULTS

## Transcriptomic Analysis of BPM and E264

We sequenced and analyzed the transcriptomes of BPM and E264 and submitted the transcriptome data to the NCBI database to obtain the sequence and annotation information of the transcriptome sequencing assembly (number: GSE147369). Relative to E264, there was no difference in the expression of six BPM homologs of *vgrG* genes (*BPM03563*, *BPM03564*, *BPM03921*, *BPM04575*, *BPM05231*, and *BPM05382*), while the expression of *BPM02934* (*vgrG2*) was upregulated, and the expression of *BPM01336* and *BPM05892* was downregulated (**Table 2**). The heatmap of *vgrG* gene expression is shown in **Figure 1**, which indicated that the *vgrG2* gene is a unique virulence factor. The VgrG protein is a needle-like structure of the T6SS and is homologous to the T4 bacteriophage cell-puncturing device, which contributes to the development of acute melioidosis (Schwarz et al., 2014). Therefore, the *vgrG2* gene was selected and subjected to further experiments in our study.

## Growth Characteristics and Antimicrobial Susceptibility Analysis of BPM, Mutant and Complemented Strains

Growth rates were plotted according to the measured $OD_{600}$ values and CFUs as described previously (Zong et al., 2019).

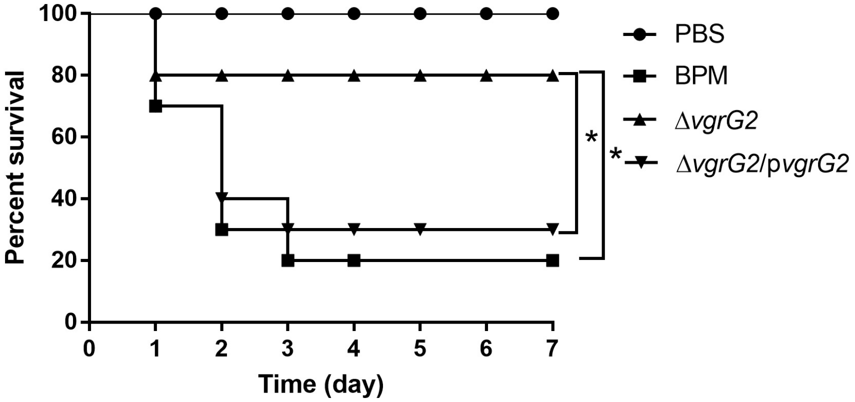


**FIGURE 2 |** Survival rate of BALB/c mice infected with BPM, mutant and complemented strains. The mortality of BALB/c mice after the intraperitoneal injection of all strains was observed over 7 days. Data points represent the percentage of BALB/c mouse survival in each group ($n$ = 10 mice per strain and 1 × $10^7$ CFU per mouse). After infection for 7 days, the survival rate of BALB/c mice infected with PBS and Δ*vgrG* was significantly lower than that of BALB/c mice infected with BPM *($P < 0.05$).

**TABLE 3 |** $LD_{50}$ of BPM, mutant and complemented strains in BALB/c mice.

| | Number of Deaths/Total | | | Mortality (%) | | |
|---|---|---|---|---|---|---|
| **Dose of Challenge CFU** | **WT** | **Δ*vgrG2*** | **Δ*vgrG2/pvgrG2*** | **WT** | **Δ*vgrG2*** | **Δ*vgrG2/pvgrG2*** |
| 8 × $10^7$ | 10/10 | 10/10 | 10/10 | 100% | 100% | 100% |
| 4 × $10^7$ | 10/10 | 10/10 | 10/10 | 100% | 100% | 100% |
| 2 × $10^7$ | 10/10 | 6/10 | 10/10 | 100% | 60% | 100% |
| 1 × $10^7$ | 8/10 | 2/10 | 7/10 | 80% | 20% | 70% |
| 5 × $10^6$ | 0/10 | 0/10 | 0/10 | 0% | 0% | 0% |
| $LD_{50}$ | | | | 8.35 × $10^6$ | 1.61 × $10^7$ | 8.87 × $10^6$ |

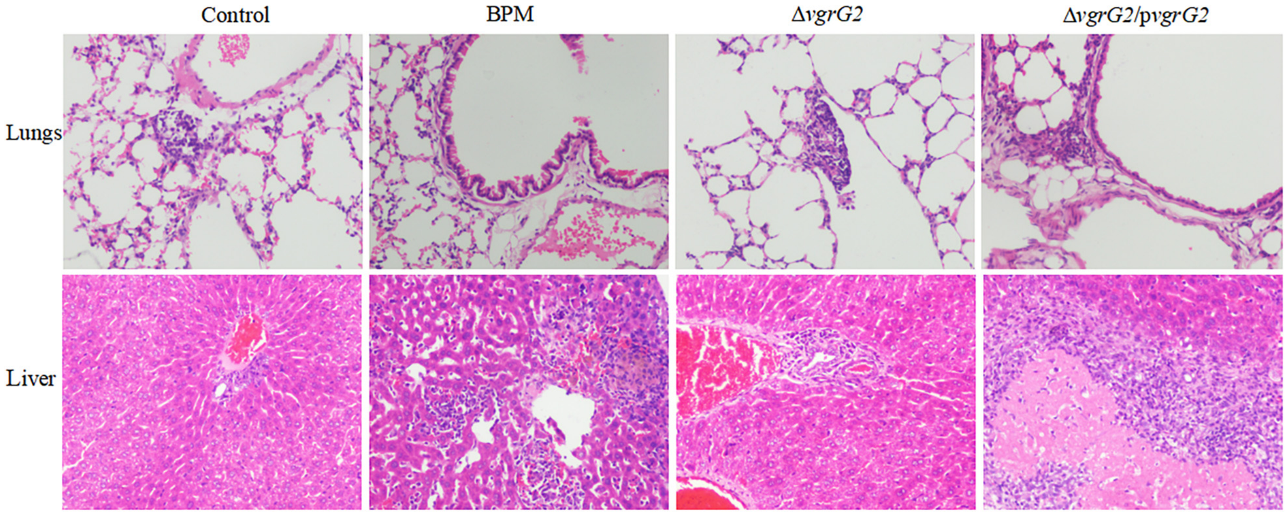


**FIGURE 3 |** Pathological characterization of lungs and liver tissues of BALB/c mice infected with BPM, mutant and complemented strains. Lungs and liver tissues of BALB/c mice infected with BPM, mutant, complemented strains and PBS (control) were prepared for light microscopy analysis and examined for differences in pathological changes (hematoxylin and eosin staining; original magnification × 200).

No growth rate difference was found when the mutant and complemented strains were compared with BPM (**Supplementary Figure 2**). The antimicrobial susceptibility results of the mutant and complemented strains related to six antibiotics were consistent with those of BPM, as shown in **Supplementary Table 1**. All of these strains were sensitive to

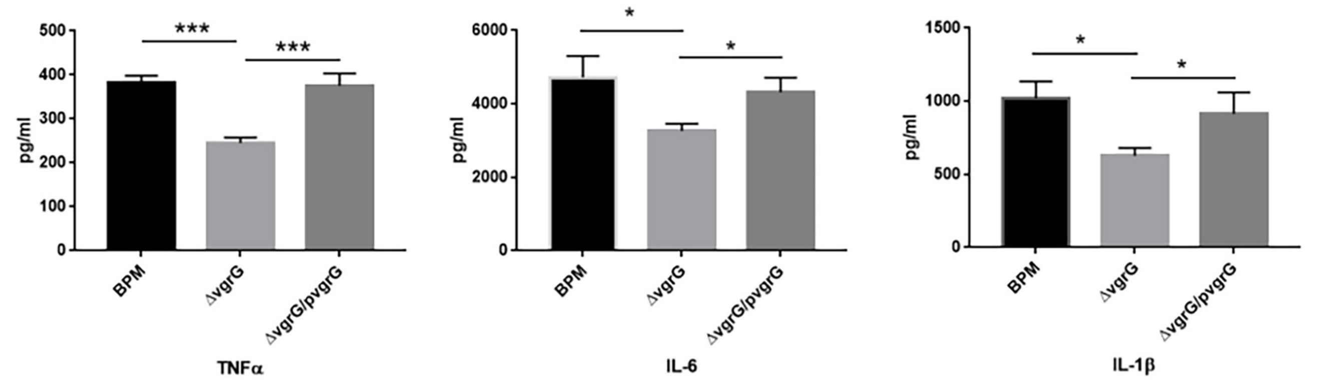


**FIGURE 4 |** Serum levels of cytokines in BLAB/c mice 16 h after infection with BPM, mutant and complemented strains. Serum IL-1β, IL-6, and TNF-α levels in BALB/c mice 16 h after infection with the BPM, mutant and complement strains. All data are from three independent experiments. Significant differences between groups are indicated: *($P$ < 0.05) and ***($P$ < 0.001).

amoxicillin/clavulanate, ceftazidime, imipenem, tetracycline, doxycycline, and trimethoprim/sulfamethoxazole.

## Survival Rate and LD$_{50}$ of BALB/c Mice Infected With BPM, Mutant and Complemented Strains

To determine whether the deletion of the *vgrG2* gene impairs the virulence of BPM, the survival rate and LD$_{50}$ of BALB/c mice infected with the BPM, mutant and complemented strains were compared. After infection for 7 days, the survival rate of BALB/c mice infected with BPM and Δ*vgrG2*/p*vgrG2* was significantly lower than that of BALB/c mice infected with Δ*vgrG2* (**Figure 2**, $^{*}P < 0.05$). The LD$_{50}$ results showed that the LD$_{50}$ of Δ*vgrG2* was $1.61 \times 10^7$ CFU (**Table 3**), which indicates low virulence. In contrast, BPM and Δ*vgrG2*/p*vgrG2* showed relatively high virulence, with LD$_{50}$ values of $8.35 \times 10^6$ CFU and $8.87 \times 10^6$ CFU, respectively (**Table 3**). Their phenotypic characteristics indicated that the *vgrG2* gene was involved in the virulence of BPM in BALB/c mice.

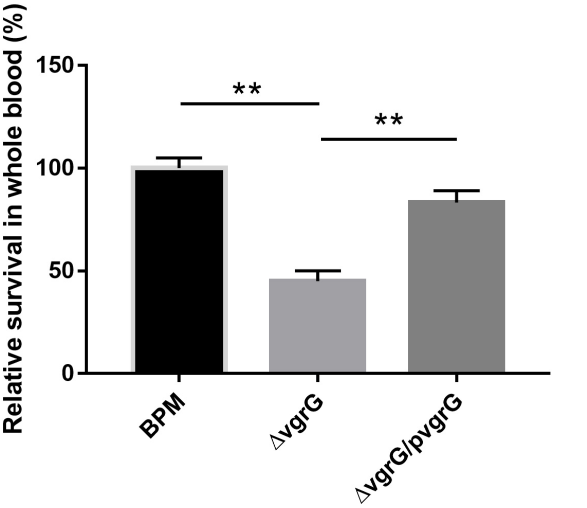


**FIGURE 5 |** Whole-blood bactericidal experiments of BPM, mutant and complemented strains. Survival rates of the BPM, mutant and complemented strains in human whole blood. The survival rates are expressed relative to those of BPM (100%). Means and SDs of three independent experiments performed in triplicate were calculated. Significant differences between groups are indicated: **($P$ < 0.01).

## Pathological Characteristics

During the first 8 h after infection, BALB/C mice infected with PBS and the three indicator strains showed no significant histopathological changes in the liver or lungs (data not shown). At 24 h after infection, different histopathological changes were observed in the lungs and livers of BALB/c mice infected with PBS and the three indicated strains. As shown in **Figure 3**, after BALB/C mice were infected with the three indicated strains, a small number of inflammatory cells infiltrated the lung tissue, central vein and convergence area, and liver tissue necrosis and partial destruction of the liver cell structure were observed.

## Deletion of *vgrG2* Decreases the Production of Inflammatory Cytokines

To determine whether the *vgrG2* gene is involved in the expression of proinflammatory cytokines, serum samples were collected from intravenously infected BALB/c mice for analysis of proinflammatory cytokines. Proinflammatory cytokines were detected in BALB/c mice 8 h after infection with the BPM, mutant and complemented strains. As shown in **Figure 4**, the production of TNF-α triggered by BPM and Δ*vgrG2*/p*vgrG2* was significantly higher than that triggered by Δ*vgrG2* (*** $P$ < 0.001). The levels of IL-1β and IL-6 induced by BPM and Δ*vgrG2*/p*vgrG2* were relatively higher than those induced by Δ*vgrG2* (* $P$ < 0.05).

## Survival Rates of BPM, Mutant and Complemented Strains in Whole Blood

To evaluate the function of the *vgrG2* gene in the evasion of innate immune responses, we measured the survival rates of the BPM, mutant and complemented strains in whole blood collected

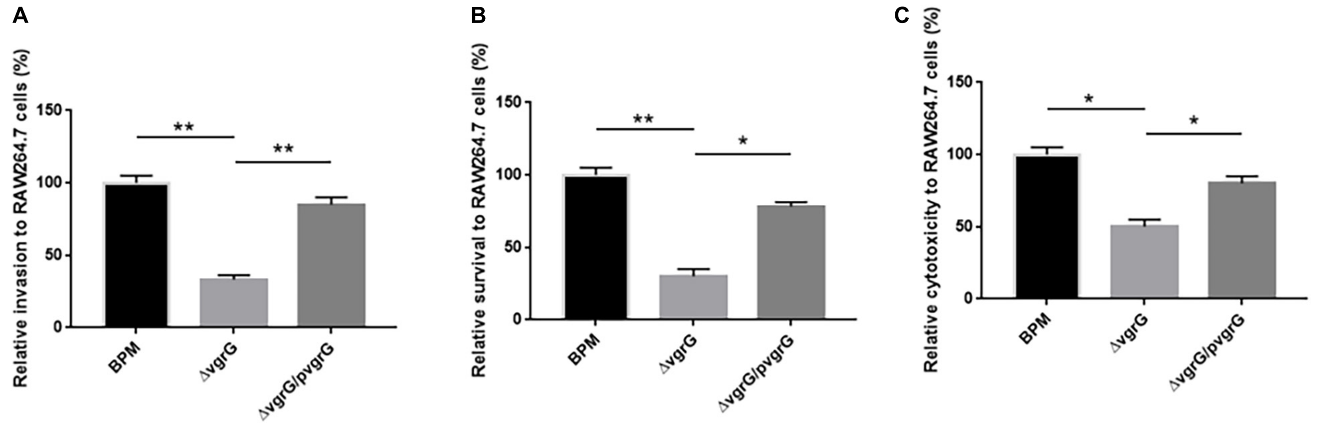


**FIGURE 6 |** Interaction between bacteria and RAW264.7 cells. **(A–C)** indicate the invasion ability, intracellular survival ability, and the cytotoxicity, respectively, of the BPM, mutant and complemented strains in RAW264.7 cells. Rates of invasion, survival and cytotoxicity are expressed relative to those of BPM (100%). The means and SDs of three independent experiments in conducted triplicate were calculated. Significant differences between groups are indicated: *($P < 0.05$) and **($P < 0.01$).

from healthy individuals. The experimental results showed that the survival rate of Δ*vgrG2* was significantly lower than those of BPM and Δ*vgrG2*/p*vgrG2* (**Figure 5**, ** $P < 0.01$).

## Interaction Between Bacteria and RAW264.7 Cells

To further investigate whether the *vgrG2* gene completely or partially impairs T6SS activity, we compared the cell invasion, intracellular survival and cytotoxicity of the mutant and complemented strains with those of BPM. The results showed that the cell invasion, intracellular survival and cytotoxicity of Δ*vgrG2* were significantly lower than those of the BPM and Δ*vgrG2*/p*vgrG2* (**Figure 6**), and no significant difference in cell invasion, intracellular survival or cytotoxicity was found between the BPM and Δ*vgrG2*/p*vgrG2* (**Figure 6**), suggesting that the deletion of *vgrG2* affected the cell invasion, intracellular survival and cytotoxicity of BPM. Since *vgrG* gene has been reported as a virulence factor of functional T6SS in *B. pseudomallei*, it can be concluded that the deletion of *vgrG2* may impair the overall activity of T6SS by affecting the assembly of T6SS in BPM.

# DISCUSSION

T6SS widely occurs in approximately 25% of all sequenced Gram-negative bacteria, including members of the genera *Vibrio*, *Pseudomonas*, *Burkholderia*, *Serratia*, *Edwardsiella*, and *Enterobacter* (Chieng et al., 2015; Gerc et al., 2015; Wood et al., 2019; Crisan and Hammer, 2020). T6SS plays an important role in pathogenicity, competition, proliferation, and cooperation (Chen et al., 2015). The T6SS is structurally, functionally, and evolutionarily related to contractile injection systems (CISs), a broad family of machines with a spring-like mechanism for delivering macromolecules into target cells (Douzi et al., 2018; Navarro-Garcia et al., 2019). A series of *Burkholderia* virulence factors, including secreted toxins, adhesins, iron acquisition systems, T6SS, and BLF1, have been reported (Bernhards et al., 2017; Lennings et al., 2018; Rust et al., 2018). Recent studies have indicated that T6SS plays an important role in the competition and pathogenicity of *Burkholderia* (Chieng et al., 2015). Our previous studies have shown that the clinical symptoms and imaging findings of patients with BPM infection are consistent with those of acute melioidosis (Chang et al., 2017). Therefore, we believe that hypervirulent *B. thailandensis* may pose a significant threat to human public health, and it is important to study the potential virulence-associated genes involved in BPM. In one of our previous studies, nine *vgrG* genes were found in the BPM genome. Further sequence analysis showed that only the *vgrG2* gene was specific to BPM. Additionally, transcriptome analysis of BPM and E264 showed that the *vgrG2* gene was strongly expressed in BPM but not in E264. Therefore, we hypothesized that the *vgrG2* gene is involved in the function of the T6SS. To test our hypothesis and describe the role of the *vgrG2* gene in BPM, a series of experiments were carried out.

To study the effect of the *vgrG2* gene on the pathogenicity of BPM, knockout mutants and complemented strains of the *vgrG2* gene were developed from BPM, and there were no significant differences in growth characteristics and antimicrobial sensitivity between them. In addition, the survival rate and $LD_{50}$ of BALB/c mice infected with these strains were compared. In animal model experiments, we found that the survival rate and $LD_{50}$ of BALB/c mice infected with Δ*vgrG2* were higher than those of BALB/c mice infected with BPM, which demonstrated that the deletion of the *vgrG2* gene significantly weakened the virulence of BPM. It has been reported that the T6SS can activate the inflammasome and cause inflammation (Aubert et al., 2016; Ratner et al., 2017; Loeven et al., 2021). To further study the function of the *vgrG2* gene in BPM pathogenesis, the serum levels of TNF-a, IL-1β and IL-6 in BALB/c mice were detected. We found that the level of IL-1β was significantly reduced after the deletion of the *vgrG2* gene. These findings indicated that the *vgrG2* gene of the

T6SS in BPM plays an important role in the pathogenicity of BPM, which was consistent with previous reports (Aubert et al., 2016). Unexpectedly, no significant differences in inflammatory cell infiltration were observed in the lungs and livers of the mice after stimulation with these different strains, particularly at 8 h after infection. The results of animal model experiments showed that the deletion of the *vgrG2* gene led to the elimination of BPM lethality and a decrease in serum cytokine levels in BALB/c mouse serum.

The evasion of innate immune responses was reported to be very important for the survival and pathogenicity of *P. aeruginosa* and *B. pseudomallei* (Gong et al., 2011; Alonso et al., 2020). In this study, our goal was to assess whether the BPM and complemented strains differ from Δ*vgrG2* in terms of their virulence and lethality. We accomplished this by measuring the survival rates of the indicated strains in whole blood. The experimental results indicated that the *vgrG2* gene may participate in the immune evasion of BPM and play a key role in the evasion of innate immune responses in whole blood. Bacterial adherence to and interaction with RAW264.7 cells are prerequisites for the induction of bacterial infection (Bruballa et al., 2020). We observed the invasion and survival abilities of the BPM, mutant and complemented strains and compared them to the abilities of RAW264.7 cells. We observed some differences when we used an MOI of 10 in our studies. Similarly, only the *vgrG2* gene affected the adherence and invasion abilities of BPM, which was consistent with the above results. In addition, previous studies have reported that the *vgrG* gene in *E. coli* and *B. pseudomallei* induced cell toxicity (Hopf et al., 2014; Cianfanelli et al., 2016). Therefore, the cytotoxicity of the BPM, mutant and complemented strains was compared, and the results were consistent with the results obtained from the examination of whole blood killing, adherence and invasion. It could be concluded that the *vgrG2* gene located within the T6SS plays a role in BPM pathogenicity, which is consistent with the hypothesis that the *vgrG2* gene is functional.

In the current study, we were unable to establish a correlation, and the virulence phenotypes of the BPM, mutant and complemented strains were similar *in vitro*, although differences in mortality were observed in the *in vivo* intravenous model of infection. Due to the wide variability of *Burkholderia* virulence properties, we strongly recommend that the selection of the tissue culture cells used in *in vitro* studies should be directly related to the cells found in the organ to which the dose will be delivered *in vivo* (Rao et al., 2020). Therefore, our study attempted to standardize the cell types used for *in vitro* and *in vivo* studies to provide a more meaningful comparison.

In conclusion, our study showed that only the *vgrG2* gene was involved in the whole blood killing of BPM, which promoted the adhesion and invasion of BPM to host cells and enhanced its pathogenicity in the host. Further studies could focus on exploring potential T6SS effectors to facilitate the development of effective antimicrobial agents for the treatment of *B. thailandensis* infection.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

The animal study was reviewed and approved by the research board of the Ethics Committee of the Third Military Medical University under permit number AMUMEC-20201085.

## AUTHOR CONTRIBUTIONS

JL, W-WH, and G-XQ performed the laboratory measurements. PJ, J-QL, W-HH, and Y-JJ made substantial contributions to the conception and design. X-RL, YX, JL, and QS participated in the experimental design and data analysis. JL drafted the manuscript. All authors read and approved the final manuscript.

## FUNDING

This study was supported by the Youth Project of the National Natural Science Foundation of China (82002113), Chongqing Medical Scientific Research Project (Joint Project of Chongqing Health Commission and Science and Technology Bureau) (No. 2022QNXM034), and Chongqing Natural Science Foundation (cstc2021jcyj-msxmX0269).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.811343/full#supplementary-material

**Supplementary Figure 1 |** Detection of *vgrG2* gene in BPM, mutant and complemented strains.

**Supplementary Figure 2 |** Growth characteristics of BPM, mutant and complemented strains.

**Supplementary Table 1 |** The antibiotic susceptibility testing results of BPM, mutant and complemented strains.

## REFERENCES

Alonso, B., Fernández-Barat, L., Di Domenico, E. G., Marín, M., Cercenado, E., Merino, I., et al. (2020). Characterization of the virulence of *Pseudomonas aeruginosa* strains causing ventilator-associated pneumonia. *BMC Infect. Dis.* 20:909. doi: 10.1186/s12879-020-05534-1

Aubert, D. F., Xu, H., Yang, J., Shi, X., Gao, W., Li, L., et al. (2016). A burkholderia type VI effector deamidates Rho GTPases to activate the pyrin inflammasome and trigger inflammation. *Cell Host Microbe* 19, 664–674. doi: 10.1016/j.chom.2016.04.004

Barnes, J. L., and Ketheesan, N. (2005). Route of infection in melioidosis. *Emerg. Infect. Dis.* 11, 638–639. doi: 10.3201/eid1104.041051

Bernhards, R. C., Cote, C. K., Amemiya, K., Waag, D. M., Klimko, C. P., Worsham, P. L., et al. (2017). Characterization of in vitro phenotypes of Burkholderia pseudomallei and Burkholderia mallei strains potentially associated with persistent infection in mice. *Arch. Microbiol.* 199, 277–301. doi: 10.1007/s00203-016-1303-8

Bobenchik, A. M., Deak, E., Hindler, J. A., Charlton, C. L., and Humphries, R. M. (2017). Performance of Vitek 2 for antimicrobial susceptibility testing of *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Stenotrophomonas* maltophilia with Vitek 2 (2009 FDA) and CLSI M100S 26th Edition Breakpoints. *J. Clin. Microbiol.* 55, 450–456. doi: 10.1128/JCM.01859-16

Brett, P. J., DeShazer, D., and Woods, D. E. (1998). *Burkholderia thailandensis* sp. nov., a Burkholderia pseudomallei-like species. *Int. J. Syst. Bacteriol.* 48, 317–320. doi: 10.1099/00207713-48-1-317

Bruballa, A. C., Shiromizu, C. M., Bernal, A. M., Pineda, G. E., Sabbione, F., Trevani, A. S., et al. (2020). Role of shiga toxins in cytotoxicity and immunomodulatory effects of *Escherichia coli* O157:H7 during host-bacterial interactions in vitro. *Toxins* 12:48. doi: 10.3390/toxins12010048

Chang, K., Luo, J., Xu, H., Li, M., Zhang, F., Li, J., et al. (2017). Human Infection with *Burkholderia thailandensis*, China, 2013. *Emerg. Infect. Dis.* 23, 1416–1418. doi: 10.3201/eid2308.170048

Chen, L., Zou, Y., She, P., and Wu, Y. (2015). Composition, function, and regulation of T6SS in *Pseudomonas aeruginosa*. *Microbiol. Res.* 172, 19–25. doi: 10.1016/j.micres.2015.01.004

Chieng, S., Mohamed, R., and Nathan, S. (2015). Transcriptome analysis of Burkholderia pseudomallei T6SS identifies Hcp1 as a potential serodiagnostic marker. *Microb. Pathog.* 79, 47–56. doi: 10.1016/j.micpath.2015.01.006

Cianfanelli, F. R., Alcoforado Diniz, J., Guo, M., De Cesare, V., Trost, M., and Coulthurst, S. J. (2016). VgrG and PAAR proteins define distinct versions of a functional type VI secretion system. *PLoS Pathog.* 12:e1005735. doi: 10.1371/journal.ppat.1005735

Crisan, C. V., and Hammer, B. K. (2020). The *Vibrio cholerae* type VI secretion system: toxins, regulators and consequences. *Environ. Microbiol.* 22, 4112–4122. doi: 10.1111/1462-2920.14976

Douzi, B., Logger, L., Spinelli, S., Blangy, S., Cambillau, C., and Cascales, E. (2018). Structure-function analysis of the C-terminal domain of the type VI secretion TssB tail sheath subunit. *J. Mol. Biol.* 430, 297–309. doi: 10.1016/j.jmb.2017.11.015

Fridman, C. M., Keppel, K., Gerlic, M., Bosis, E., and Salomon, D. (2020). A comparative genomics methodology reveals a widespread family of membrane-disrupting T6SS effectors. *Nat. Commun.* 11:1085. doi: 10.1038/s41467-020-14951-4

Garcia, E. C. (2017). *Burkholderia thailandensis*: genetic Manipulation. *Curr. Protoc. Microbiol.* 45, 4C.2.1–4C.2.15. doi: 10.1002/cpmc.27

Gassiep, I., Armstrong, M., and Norton, R. (2020). Human melioidosis. *Clin. Microbiol. Rev.* 33, e00006–19. doi: 10.1128/CMR.00006-19

Gee, J. E., Elrod, M. G., Gulvik, C. A., Haselow, D. T., Waters, C., Liu, L., et al. (2018). *Burkholderia thailandensis* isolated from infected wound, Arkansas, USA. *Emerg. Infect. Dis.* 24, 2091–2094. doi: 10.3201/eid2411.180821

Gerc, A. J., Diepold, A., Trunk, K., Porter, M., Rickman, C., Armitage, J. P., et al. (2015). Visualization of the serratia type VI secretion system reveals unprovoked attacks and dynamic assembly. *Cell Rep.* 12, 2131–2142. doi: 10.1016/j.celrep.2015.08.053

Ginther, J. L., Mayo, M., Warrington, S. D., Kaestli, M., Mullins, T., Wagner, D. M., et al. (2015). Identification of burkholderia pseudomallei near-neighbor species in the northern territory of Australia. *PLoS Negl. Trop. Dis.* 9:e0003892. doi: 10.1371/journal.pntd.0003892

Glass, M. B., Gee, J. E., Steigerwalt, A. G., Cavuoti, D., Barton, T., Hardy, R. D., et al. (2006). Pneumonia and septicemia caused by *Burkholderia thailandensis* in the United States. *J. Clin. Microbiol.* 44, 4601–4604. doi: 10.1128/JCM.01585-06

Gong, L., Cullinane, M., Treerat, P., Ramm, G., Prescott, M., Adler, B., et al. (2011). The Burkholderia pseudomallei type III secretion system and BopA are required for evasion of LC3-associated phagocytosis. *PLoS One* 6:e17852. doi: 10.1371/journal.pone.0017852

Hopf, V., Göhler, A., Eske-Pogodda, K., Bast, A., Steinmetz, I., and Breitbach, K. (2014). BPSS1504, a cluster 1 type VI secretion gene, is involved in intracellular survival and virulence of Burkholderia pseudomallei. *Infect. Immun.* 82, 2006–2015. doi: 10.1128/IAI.01544-14

Hsieh, P. F., Lu, Y. R., Lin, T. L., Lai, L. Y., and Wang, J. T. (2019). *Klebsiella pneumoniae* Type VI secretion system contributes to bacterial competition, cell invasion, type-1 fimbriae expression, and in vivo colonization. *J. Infect. Dis.* 219, 637–647. doi: 10.1093/infdis/jiy534

Jiang, X., Yang, Y., Zhou, J., Zhu, L., Gu, Y., Zhang, X., et al. (2016). Roles of the putative type IV-like secretion system key component VirD4 and PrsA in pathogenesis of streptococcus suis type 2. *Front. Cell. Infect. Microbiol.* 6:172. doi: 10.3389/fcimb.2016.00172

Kovacs-Simon, A., Hemsley, C. M., Scott, A. E., Prior, J. L., and Titball, R. W. (2019). *Burkholderia thailandensis* strain E555 is a surrogate for the investigation of Burkholderia pseudomallei replication and survival in macrophages. *BMC Microbiol.* 19:97. doi: 10.1186/s12866-019-1469-8

Lennings, J., West, T. E., and Schwarz, S. (2018). The Burkholderia type VI secretion system 5: composition, regulation and role in virulence. *Front. Microbiol.* 9:3339. doi: 10.3389/fmicb.2018.03339

Lertpatanasuwan, N., Sermsri, K., Petkaseam, A., Trakulsomboon, S., Thamlikitkul, V., and Suputtamongkol, Y. (1999). Arabinose-positive Burkholderia pseudomallei infection in humans: case report. *Clin. Infect. Dis.* 28, 927–928. doi: 10.1086/517253

Lewis, E., Kilgore, P. B., Mott, T. M., Pradenas, G. A., and Torres, A. G. (2017). Comparing in vitro and in vivo virulence phenotypes of Burkholderia pseudomallei type G strains. *PLoS One* 12:e0175983. doi: 10.1371/journal.pone.0175983

Loeven, N. A., Perault, A. I., Cotter, P. A., Hodges, C. A., Schwartzman, J. D., Hampton, T. H., et al. (2021). The Burkholderia cenocepacia type VI secretion system effector TecA is a virulence factor in mouse models of lung infection. *mBio* 12:e0209821. doi: 10.1128/mBio.02098-21

Luo, P., He, X., Liu, Q., and Hu, C. (2015). Developing universal genetic tools for rapid and efficient deletion mutation in vibrio species based on suicide T-vectors carrying a novel counterselectable marker, vmi480. *PLoS One* 10:e0144465. doi: 10.1371/journal.pone.0144465

Navarro-Garcia, F., Ruiz-Perez, F., Cataldi, Á., and Larzábal, M. (2019). Type VI secretion system in pathogenic *Escherichia coli*: structure, role in virulence, and acquisition. *Front. Microbiol.* 10:1965. doi: 10.3389/fmicb.2019.01965

Ngamdee, W., Tandhavanant, S., Wikraiphat, C., Reamtong, O., Wuthiekanun, V., Salje, J., et al. (2015). Competition between Burkholderia pseudomallei and B. thailandensis. *BMC Microbiol.* 15:56. doi: 10.1186/s12866-015-0395-7

Pijuan, J., Barceló, C., Moreno, D. F., Maiques, O., Sisó, P., Marti, R. M., et al. (2019). In vitro cell migration, invasion, and adhesion assays: from cell imaging to data analysis. *Front. Cell Dev. Biol.* 7:107. doi: 10.3389/fcell.2019.00107

Rao, C., Mao, C., Xia, Y., Zhang, M., Hu, Z., Yuan, S., et al. (2020). Transcriptome analysis reveals unfolded protein response was induced during the early stage of burkholderia pseudomallei infection in A549 cells. *Front. Genet.* 11:585203. doi: 10.3389/fgene.2020.585203

Ratner, D., Orning, M. P., and Lien, E. (2017). Bacterial secretion systems and regulation of inflammasome activation. *J. Leukoc. Biol.* 101, 165–181. doi: 10.1189/jlb.4MR0716-330R

Rust, A., Shah, S., Hautbergue, G. M., and Davletov, B. (2018). Burkholderia lethal factor 1, a novel anti-cancer toxin, demonstrates selective cytotoxicity in MYCN-amplified neuroblastoma cells. *Toxins* 10:261. doi: 10.3390/toxins10070261

Schwarz, S., Singh, P., Robertson, J. D., LeRoux, M., Skerrett, S. J., Goodlett, D. R., et al. (2014). VgrG-5 is a Burkholderia type VI secretion system-exported protein required for multinucleated giant cell formation and virulence. *Infect. Immun.* 82, 1445–1452. doi: 10.1128/IAI.01368-13

Tang, X., Li, G., Shi, L., Su, F., Qian, M., Liu, Z., et al. (2021). Combined intermittent fasting and ERK inhibition enhance the anti-tumor effects of chemotherapy via the GSK3β-SIRT7 axis. *Nat. Commun.* 12:5058. doi: 10.1038/s41467-021-25274-3

Tang, Y., Deng, J., Zhang, J., Zhong, X., Qiu, Y., Zhang, H., et al. (2018). Epidemiological and clinical features of melioidosis: a report of seven cases from Southern Inland China. *Am. J. Trop. Med. Hyg.* 98, 1296–1299. doi: 10.4269/ajtmh.17-0128

Wiersinga, W. J., Currie, B. J., and Peacock, S. J. (2012). Melioidosis. *N. Engl. J. Med.* 367, 1035–1044. doi: 10.1056/NEJMra1204699

Wood, T. E., Howard, S. A., Förster, A., Nolan, L. M., Manoli, E., Bullen, N. P., et al. (2019). The *Pseudomonas aeruginosa* T6SS delivers a periplasmic toxin that disrupts bacterial cell morphology. *Cell Rep.* 29, 187–201.e7. doi: 10.1016/j.celrep.2019.08.094

Zhao, Y., Liu, G., Li, S., Wang, M., Song, J., Wang, J., et al. (2011). Role of a type IV-like secretion system of Streptococcus suis 2 in the development of streptococcal toxic shock syndrome. *J. Infect. Dis.* 204, 274–281. doi: 10.1093/infdis/jir261

Zhu, Q., Chen, X., Liu, Y., Wang, R., Chen, J., and Chen, Y. (2020). Virulence, antimicrobial susceptibility, molecular and epidemiological characteristics of a new serotype of vibrio parahaemolyticus from diarrhea patients. *Front. Microbiol.* 11:2025. doi: 10.3389/fmicb.2020.02025

Zhu, Y., Chen, P., Bao, Y., Men, Y., Zeng, Y., Yang, J., et al. (2016). Complete genome sequence and transcriptomic analysis of a novel marine strain Bacillus weihaiensis reveals the mechanism of brown algae degradation. *Sci. Rep.* 6:38248. doi: 10.1038/srep38248

Zong, B., Zhang, Y., Wang, X., Liu, M., Zhang, T., Zhu, Y., et al. (2019). Characterization of multiple type-VI secretion system (T6SS) VgrG proteins in the pathogenicity and antibacterial activity of porcine extra-intestinal pathogenic *Escherichia coli*. *Virulence* 10, 118–132. doi: 10.1080/21505594.2019.1573491

**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.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 May 2022
doi: 10.3389/fmicb.2022.892333

# Global Phylogeny of *Mycobacterium avium* and Identification of Mutation Hotspots During Niche Adaptation

*Rachel Mizzi[1], Karren M. Plain[1,2]*, Richard Whittington[1] and Verlaine J. Timms[3]*

*[1]Farm Animal Health, School of Veterinary Science, Faculty of Science, The University of Sydney, Camden, NSW, Australia, [2]Microbiology and Parasitology Research, Elizabeth Macarthur Agricultural Institute, Menangle, NSW, Australia, [3]Neilan Laboratory of Microbial and Molecular Diversity, College of Engineering, Science and Environment, The University of Newcastle, Newcastle, NSW, Australia*

*Mycobacterium avium* is separated into four subspecies: *M. avium* subspecies *avium* (MAA), *M. avium* subspecies *silvaticum* (MAS), *M. avium* subspecies *hominissuis* (MAH), and *M. avium* subspecies *paratuberculosis* (MAP). Understanding the mechanisms of host and tissue adaptation leading to their clinical significance is vital to reduce the economic, welfare, and public health concerns associated with diseases they may cause in humans and animals. Despite substantial phenotypic diversity, the subspecies nomenclature is controversial due to high genetic similarity. Consequently, a set of 1,230 *M. avium* genomes was used to generate a phylogeny, investigate SNP hotspots, and identify subspecies-specific genes. Phylogeny reiterated the findings from previous work and established that *Mycobacterium avium* is a species made up of one highly diverse subspecies, known as MAH, and at least two clonal pathogens, named MAA and MAP. Pan-genomes identified coding sequences unique to each subspecies, and in conjunction with a mapping approach, mutation hotspot regions were revealed compared to the reference genomes for MAA, MAH, and MAP. These subspecies-specific genes may serve as valuable biomarkers, providing a deeper understanding of genetic differences between *M. avium* subspecies and the virulence mechanisms of mycobacteria. Furthermore, SNP analysis demonstrated common regions between subspecies that have undergone extensive mutations during niche adaptation. The findings provide insights into host and tissue specificity of this genetically conserved but phenotypically diverse species, with the potential to provide new diagnostic targets and epidemiological and therapeutic advances.

**Keywords:** ***Mycobacterium avium*, *hominissuis*, *paratuberculosis*, *silvaticum*, comparative genomics, mutation hotspot**

OPEN ACCESS

***Edited by:***
*Zhan Zhou,*
*Zhejiang University, China*

***Reviewed by:***
*Marcel Behr,*
*McGill University, Canada*
*John Bannantine,*
*National Animal Disease Center,*
*Agricultural Research Service*
*(USDA), United States*
*Leonardo A. Sechi,*
*University of Sassari, Italy*

***Correspondence:**
*Karren M. Plain*
*karren.plain@sydney.edu.au*

***Specialty section:***
*This article was submitted to*
*Evolutionary and Genomic*
*Microbiology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:*** *08 March 2022*
***Accepted:*** *06 April 2022*
***Published:*** *06 May 2022*

***Citation:***
*Mizzi R, Plain KM, Whittington R and Timms VJ (2022) Global Phylogeny of Mycobacterium avium and Identification of Mutation Hotspots During Niche Adaptation.*
*Front. Microbiol. 13:892333.*
*doi: 10.3389/fmicb.2022.892333*

## INTRODUCTION

The *Mycobacterium avium* complex (MAC) is a group of slow-growing (>1 week to form visible colonies during culture) non-tuberculosis mycobacteria (NTM). A recent definition of these species was described by van Ingen et al. (2018). They characterized MAC species by a sequence identity of >99.4% for the full *16S* rRNA gene, >97.3% for *hsp65*, and > 94.4% for *rpoB* region

V for reference stains *Mycobacterium intracellulare* ATCC 13950 (Kim et al., 2012) or *Mycobacterium avium* ATCC 25291 (Goethe et al., 2020). According to this definition, the MAC contains 12 species: *Mycobacterium avium* (Thorel et al., 1990), *Mycobacterium intracellulare* (Wayne et al., 1993), *Mycobacterium chimaera* (Tortoli et al., 2004), *Mycobacterium colombiense* (Murcia et al., 2006), *Mycobacterium arosiense* (Bang et al., 2008), *Mycobacterium vulneris* (van Ingen et al., 2009), *Mycobacterium bouchedurhonense*, *Mycobacterium timonense*, *Mycobacterium marseillense* (Ben Salah et al., 2009), *Mycobacterium yongonense* (Kim et al., 2013), *Mycobacterium paraintracellulare* (Lee et al., 2016a), and *Mycobacterium lepraemurium*.

*Mycobacterium avium* has been separated into four subspecies: *M. avium* subspecies *avium* (MAA), *M. avium* subspecies *silvaticum* (MAS), *M. avium* subspecies *hominissuis* (MAH), and *M. avium* subspecies *paratuberculosis* (MAP; **Figure 1**). Understanding the clinical significance and mechanisms of host and tissue adaptation of these subspecies is vital to reduce economic, welfare, and public health concerns associated with the diseases they cause in humans and animals. Despite the obvious phenotypic diversity of these subspecies, their nomenclature is controversial due to their high degree of genetic similarity. Furthermore, little is known about the biological reasons for different *M. avium* subspecies to infect and survive in different host and tissue niches.

Early descriptions and infection trials demonstrated differences in the pathogenicity and host range of ruminant and avian mycobacterial isolates, leading to the hypothesis that there were several *M. avium* subspecies (Collins et al., 1985). In 1990, three *M. avium* subspecies were recognized (Thorel et al., 1990), with the former species *M. paratuberculosis* being included as a subspecies and named *M. avium* subsp. *paratuberculosis* (MAP). Differences between human and porcine isolates and avian strains were identified using molecular methods, and this led to the nomination of MAH for these *M. avium* isolates (Mijs et al., 2002). MAP is arguably the most studied pathogen in the *M. avium* complex and is the causative agent of paratuberculosis or Johne's disease (JD), a chronic gastroenteritis that predominately affects ruminants. This pathogen has also been implicated in the pathogenesis of Crohn's disease (Bach, 2015; Waddell et al., 2015; Timms et al., 2016), type 1 diabetes, and multiple sclerosis (Eslami et al., 2019; Ekundayo et al., 2022) in humans. The tissue tropism of MAP in ruminants is the gastrointestinal tract, specifically the ileum, though disseminated infection to other organs and tissues occurs as the disease progresses. This primary site of infection is unique to this subspecies; other subspecies tend to preferentially infect the respiratory tract or are acute disseminated infections with no specific tissue preference.

*Mycobacterium avium* subspecies *avium* and MAS typically cause respiratory disease in avian species, with the latter almost exclusively restricted to wood pigeons (*Columba palumbus*). Tuberculosis-like respiratory disease caused by MAA in avian species is a common disease that can be economically important due to high mortalities and has welfare concerns, but it is less often reported in the literature than JD in ruminants (Moravkova et al., 2013; Salamatian et al., 2020). MAS are the least reported of the subspecies, and little is known about this mycobacterium due to the relatively small number of isolates available for study.

*Mycobacterium avium* infections reported in humans and swine are typically caused by MAH. In humans, cases typically present as pulmonary disease in immunocompetent individuals, peripheral lymphadenopathy in children, or disseminated infection in immunocompromised patients (Slany et al., 2016). Cases in immunocompetent individuals are particularly concerning due to the high prevalence of antimicrobial resistance of among *M. avium* isolates, particularly MAH (Brown-Elliott et al., 2012; Wang et al., 2021). In swine, mesenteric, cranial, or cervical lymph node lesions are the most common clinical

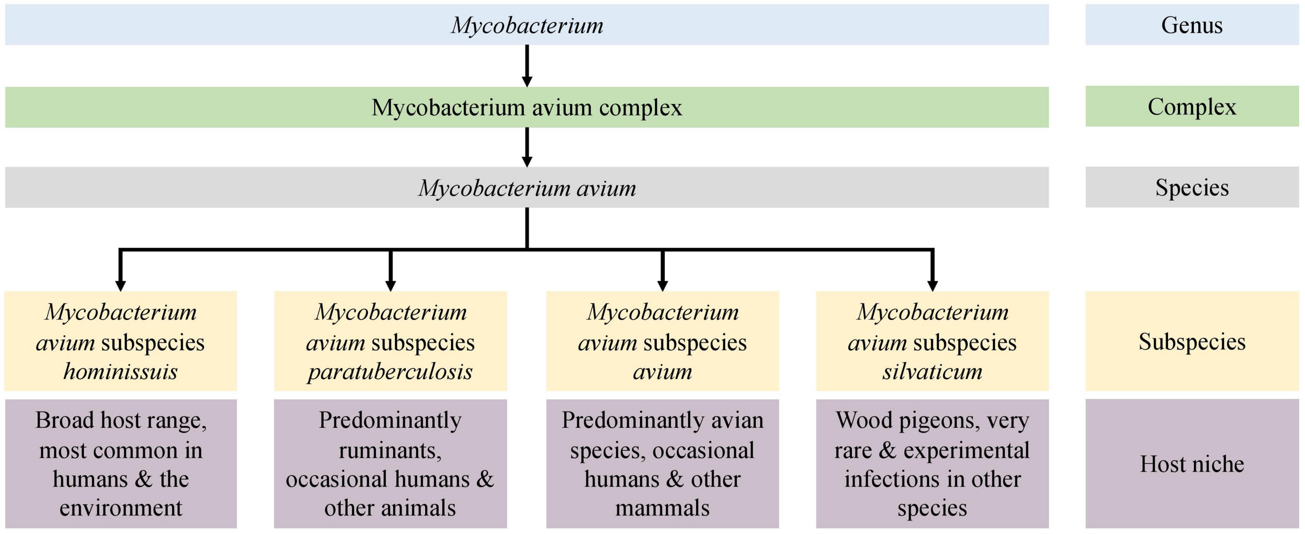

Mycobacterium
Genus
Mycobacterium avium complex
Complex
Mycobacterium avium
Species
Mycobacterium avium subspecies hominissuis
Mycobacterium avium subspecies paratuberculosis
Mycobacterium avium subspecies avium
Mycobacterium avium subspecies silvaticum
Subspecies
Broad host range, most common in humans & the environment
Predominantly ruminants, occasional humans & other animals
Predominantly avian species, occasional humans & other mammals
Wood pigeons, very rare & experimental infections in other species
Host niche

**FIGURE 1** | An overview of the *Mycobacterium avium* subspecies.

presentations, and often, no ante-mortem clinical signs are apparent (Slany et al., 2016).

Differentiation of *M. avium* to the subspecies level in clinical practice is hindered by the need for specialized methods typically confined to research only. Common typing techniques include restriction fragment length polymorphism (RFLP) analysis utilizing various insertion sequences (IS; Mijs et al., 2002; Johansen et al., 2005; Moravkova et al., 2008; Rindi and Garzelli, 2014) or variable number tandem repeats (VNTRs) typing using mycobacterial interspersed repetitive units (MIRUs; Radomski et al., 2010; Rindi and Garzelli, 2014). Ambiguities can arise from these typing techniques as some IS elements share high sequence identity (Johansen et al., 2005) and VNTR–MIRU discrimination may not be sufficient to distinguish some isolates (Pate et al., 2011).

Evidence for the close relationship between MAA and MAS is abundant; however, the relationships between all subspecies have not been widely studied (Turenne et al., 2007; Paustian et al., 2008; Radomski et al., 2010). The true genetic diversity present within each subspecies of *M. avium* is another knowledge gap. Previous studies have focused on type strains (Bannantine et al., 2012; Möbius et al., 2015); a limited number of genes (Turenne et al., 2008) or a small number of isolates (Wibberg et al., 2020; Bannantine et al., 2020a,b). Microarray technology has revealed several large sequence polymorphisms between avian (MAA) and ruminant (MAP) isolates (Paustian et al., 2005, 2008). However, this technique uses a single reference strain to compare against other isolates. Limited conclusions can be drawn for isolates that were not directly compared to the reference strain. Furthermore, genomic regions that are absent from the reference strain but present in other isolates may not be recognized. A recent study utilized 29 closed genomes and discovered several genes that were subspecies-specific. However, this investigation was limited by a small dataset.

Whole-genome comparisons intuitively would allow accurate and comprehensive comparison of isolates; however, such methods can also be problematic for subspecies delimitation when a small number of isolates are used to describe taxa. The 70% DNA–DNA hybridization (DDH; Meier-Kolthoff et al., 2013) or 97% average nucleotide identity (ANI; Lee et al., 2016b) cut offs that have traditionally been used for species delimitation fail to distinguish between *M. avium* subspecies and other closely related MAC mycobacteria (Riojas et al., 2018; Tortoli et al., 2019). The concern with these methods is their reliance on a single strain type to represent a species. This can create complications in downstream analysis when intraspecies diversity makes it difficult to classify new isolates of unknown species. Furthermore, relative to many other bacterial genera, mycobacteria are a genetically homogenous group, yet they have diverse lifestyles and growth characteristics and exist in a broad range of niches. This indicates that the variability that does exist is biologically significant, and an accurate resolution of this variability is required. Consequently, DDH and ANI cut off values alone may not to be appropriate for the definition of mycobacterial species.

The understanding of *M. tuberculosis* diversity and lineages enables efficient outbreak tracing (Gardy et al., 2011) and informs epidemiologists to enable identification of the source of an outbreak and formulation of optimal control measures (Didelot et al., 2016). Arguably, identification of *M. avium* pathogens to the subspecies level is also crucial for understanding their significance and to perform epidemiological studies. However, relatively recent taxonomic studies concluded that the subspecies should be removed from *M. avium* taxonomy as the threshold for subspecies demarcation is not reached (Riojas et al., 2018; Tortoli et al., 2019). This conclusion was based on results from single type strains of MAP, MAA, and MAH. Regardless of the nomenclature, significant biological and phenotypic differences between the subspecies of *M. avium* were recognized by Thorel and Mijs (Thorel et al., 1990; Mijs et al., 2002) in niche adaptation, host preference, and growth characteristics (**Figure 1**). The availability of whole-genome sequencing (WGS) and the expansion of public genomic databases provide an opportunity to study many *M. avium* genomes and to make recommendations based on comprehensive analysis.

In this large-scale study, publicly available and 73 newly sequenced *M. avium* genomes were analyzed to compare subspecies clusters using pan-genome and SNPs analysis approaches, to identify subspecies-specific genes and mutation hotspots. A panel of subspecies-specific genes were identified that may serve as valuable biomarkers, and the SNP hotspot analysis demonstrated common regions between subspecies that have undergone extensive mutations during niche adaptation. Together, these outcomes will inform epidemiological analysis, lead to better disease control in animals and so reduce the chance of spillover into humans.

## MATERIALS AND METHODS

### Isolate and Metadata Collation

Publicly available WGS data of *M. avium* isolates were sourced from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and the NCBI GenBank. Searches on public databases were undertaken on the 2/3/2021 for "*Mycobacterium avium*" and in the SRA database; filters for public, DNA, genome, paired, and Illumina were used. No raw reads were available from GenBank so assemblies were used. Any SRA isolate labelled as anything other than *Mycobacterium avium* or one of the subspecies (*silvaticum*, *paratuberculosis*, *avium*, or *hominissuis*) was removed from the dataset. Available public metadata are summarized in the excel spreadsheet in **Supplementary Material 1**. Additionally, 10 isolates were sequenced by the Mycobacterium Reference Laboratory, Westmead Hospital, NSW, 48 animal derived isolates were sequenced by the Farm Animal Health Group, University of Sydney, and 15 archival isolates of *M. avium* provided by the Mycobacterial Reference Laboratory Queensland were sequenced for this study. The archival isolates were recognized strains of the International Working Group on Mycobacterial Taxonomy (IWGMT; Wayne et al., 1993) and represented most serovars of *M. avium*: serovar 1 (Strains 17 and 1), serovar 2 (Strains 19, 55, and 60), serovar 3 (Strains 26 and 38), serovar 4 (Strains 54 and 62), serovar 5 (Strain 23), serovar 8 (Strain

29), serovar 9 (Strains 18 and 28), serovar 10 (Strain 49), and serovar 11 (Strain 31; Wayne et al., 1993). These serovar isolates were selected to ensure a set of representative serovars that were included in this investigation. Their details are summarized in the excel spreadsheet in **Supplementary Material 1**.

## Culture and DNA Extraction

To ensure that a high quantity and quality of DNA was extracted, a lengthy culture and DNA extraction procedure were undertaken for isolates sequenced by the University of Sydney as previously described (Mizzi et al., 2021). The process involved mechanical and enzymatic cell wall digestion, followed by a combination of hexadecyltrimethylammonium bromide/saline (CTAB/NaCl) and phenol–chloroform extraction.

## Quality Control and Assembly

All fastq files were trimmed using Trimmomatic (version 0.36, RRID:SCR_011848; Bolger et al., 2014) with options set to -phred33, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36. Any genome with a g-zipped fastq file (forward, reverse, or both) with less than 50,000,000 bytes after trimming was discarded. Reads were assembled with SPAdes (version 3.12.0, RRID:SCR_000131; Bankevich et al., 2012) using the default k-mer size testing options. To improve the assemblies, the Bayes–Hammer read correction, and careful option for post-assembly Burrows Wheeler Aligner mismatch correction (Li and Durbin, 2009) were used. Quality assessment of the newly created assemblies and those obtained from GenBank were done with QUAST (version 5.0.2, RRID:SCR_001228; Gurevich et al., 2013). Assemblies with a GC% of less than 68%, number of contigs >500, or a total length outside of 4.5–6.2 megabases were removed from the study. A summary of the bioinformatics methods is depicted in **Figure 2**.

## Pan-Genome Analysis

Genome annotation was undertaken with Prokka (version 1.14.5, RRID:SCR_01473; Seemann, 2014) with the minimum contig length set to 500 base pairs. The Panaroo (version 1.2.3, RRID:SCR_021090; Tonkin-Hill et al., 2020) pan-genome pipeline was used for pan-genome analysis using the parameters for strict clean mode to remove contaminants and a sequence identity threshold of 0.75. Within the total number of genes identified by Panaroo software in the pan-genome analysis, there were several categories of genes distinguished by the number of isolates that contained a given gene. Core genes are those present in 99%–100% of isolates. The accessory genome, or portion not present in all isolates, is split into the soft core genes that are present in ≤95 to <99% of genomes, shell genes that are present in 15% ≤ to <95%, and cloud genes, which are present in <15% of strains.

## SNP Analysis

All genomes were mapped to one closed reference strain for each subspecies except for MAS as there was no closed reference genome available. These included the K10 strain

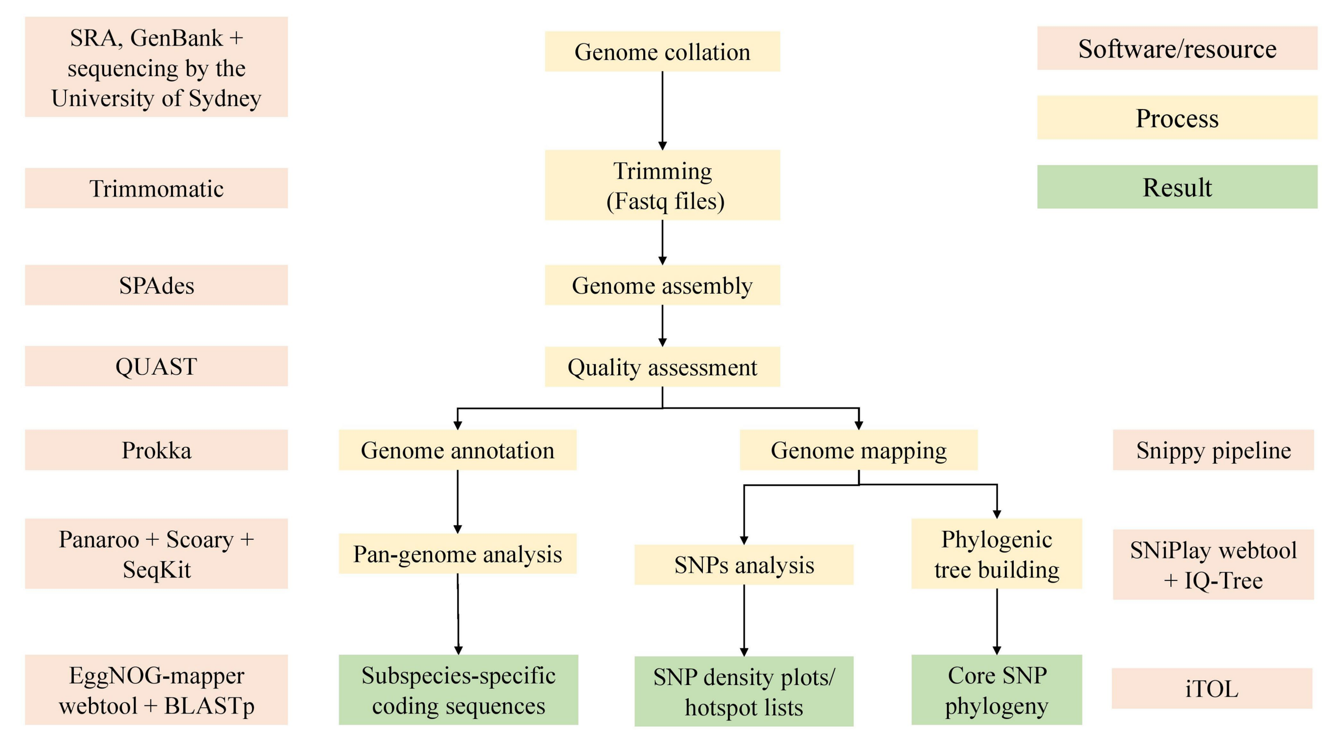


**FIGURE 2 |** An overview of the bioinformatic methods used in this study. Software used included Trimmomatic (version 0.36; Bolger et al., 2014), SPAdes (Bankevich et al., 2012), QUAST (Gurevich et al., 2013), Prokka (Seemann, 2014), Panaroo (Tonkin-Hill et al., 2020), Snippy pipeline (Seemann, 2019), SNiPlay (Dereeper et al., 2011), EggNOG (Huerta-Cepas et al., 2017), IQ-Tree (Nguyen et al., 2015), and iTol (Letunic and Bork, 2019).

for MAP, strain MAH104 for MAH and the Chester (DSM44156) strain for MAA. Summary statistics of the reference genomes is available in the supplementary tables (**Supplementary Material 4**). The Snippy pipeline (version 3.1; available at https://github.com/tseemann/snippy) was used with the default settings. Genome assemblies were used as the input given raw read files were not available for genomes downloaded from GenBank. The core.txt output file was viewed to identify the percentage of artificial reads mapping to the K10 reference genome. The core SNP alignment output file from Snippy was used by IQ-Tree to create a tree based on SNPs in the core genome. In the initial preliminary tree, the *Mycobacterium intracellulare* type strain ATCC13950 and *Mycobacterium intracellulare* subspecies chimaera type strain DSM44623 were added to the dataset. Any genomes that clustered with ATCC13950 or DSM44623 were removed from the study. The output using the MAP K10 reference genome was used as the final tree. Other tree outputs using the MAH and MAA reference genome are available in the supplementary figures (**Supplementary Material 2**).

To identify potential SNP hotspots in the subspecies, we assessed the SNP density throughout the dataset. This was done by using the location of high-quality core SNPs present in the combined VCF output from Snippy. This file was used as an input for the SNiPlay variant density viewer webtool (Dereeper et al., 2011). A sliding window of 10,000 bp was used in SNiPlay to produce SNP density plots. SNP hotspots were defined as a region that had a SNP density greater than four SDs from the mean SNP density across the reference genome. The 10,000 bp regions that displayed the highest number of SNPs from each dataset to each reference genome were extracted from the reference genome using bedtools. Hotspot regions were annotated with Prokka without the minimum contig length restriction, and the resulting protein fasta files were uploaded to the EggNOG functional annotation mapper webtool (version 2; Huerta-Cepas et al., 2017) with the taxonomic scope set to *Mycobacteriaceae* for further annotation.

## Phylogenetic Analyses

IQ-Tree (version 1.6.7, RRID:SCR_017254; Nguyen et al., 2015) was used to generate trees using the GTR + I + G model. Trees were visualized and annotated in iTol (RRID:SCR_018174; Letunic and Bork, 2019). Initially, the type strains for *Mycobacterium intracellulare* and *Mycobacterium chimaera* were included in the dataset to determine whether any genomes in the dataset were incorrectly labelled as *Mycobacterium avium*. Genomes that clustered with *M. intracellulare* or *M. chimaera* were very distant from majority of genomes in the study and removed from the dataset ($n = 255$, data not shown). Isolates were grouped into subspecies based on their clustering in the tree. Since only two typed isolates were available for the MAS subspecies, bird-type subspecies MAA and MAS were grouped together in this investigation and are referred to collectively as MAA/S. The phylogeny was rooted at the midpoint in iTol for ease of readability.

## Querying the Pan-Genome

Genes of interest for this study were those present in all isolates from one subspecies and absent in other subspecies. Genome-wide associations between genes identified in the pan-genome and subspecies were achieved with Scoary (RRID:SCR_021087; Brynildsrud et al., 2016) and validated with BLASTp. An association from Scoary was considered significant where a gene had a corrected value of *p* (Benjamini-Hochberg) of less than 0.05. Sequences of the genes of interest were extracted from the pan-genome reference output from Panaroo using seqkit (version 0.10.1; Shen et al., 2016). Coding sequences were converted from DNA sequences to protein sequences using Prokka. BLASTp was used to confirm the subspecies-specificity of each gene of interest. A gene was considered subspecies-specific if it had a matching identity of 97% or higher, *e*-value less than 0.000001, bitscore of 100, or greater and coverage of 95% of the full gene length or greater. Coding sequences of the markers identified by Scoary and validated by BLASTp were uploaded to EggNOG (available at http://eggnog-mapper.embl.de/, RRID:SCR_002456) with the taxonomic scope set to *Mycobacteriaceae*. To search for markers of pathogenicity, this approach was also used to compare the clonal pathogenic subspecies (MAA, MAS, and MAP) to the opportunistic pathogen MAH. A similar approach was taken to identify genes associated with respiratory or gastrointestinal tissue trophism, where MAP or not MAP was the trait of interest.

# RESULTS

## Whole-Genome Sequencing and Assembly

Overall, the quality of published genomes and assemblies created from public raw data was highly variable. Many genomes that were accessible from the Sequence Read Archive (SRA) and GenBank were of undesirable quality and did not meet quality thresholds outline in the methods section (very high number of contigs, abnormally large or small file size or genome length, or GC% very low indicating contamination) for this investigation and thus were not included in the study.

At the time of searching (02/08/2021), 3,167 genomes were available on Sequence Read Archive (SRA) and 216 in GenBank. Of the total 3,383 available, 60 were removed as they were labeled as something other than *mycobacterium avium* such as *M. chimaera* or *M. intracellulare* in the metadata downloaded from the SRA. Post-trimming, 250 samples had one or more compressed Fastq files smaller than 50,000,000 bytes, 12 genomes failed to assemble, 226 samples had a GC% <68, 587 had more than 500 contigs, and 121 had a length outside of 4.5–5.3 megabases. Some samples did not meet multiple QUAST criteria.

During preliminary SNP analysis, the Snippy pipeline demonstrated that the MAH reference genome MAH104, MAA reference DSM44156 and MAP sheep strain Telford had 93.61, 93.78, and 98.9 percent of artificial reads map to the K10 MAP reference genome. Any sample that had less than 93% of artificial reads map to K10 was discarded ($n = 624$). An exception was made for the silvaticum reference genome, which had 87.25% of artificial reads map to K10. Once quality thresholds were met, an additional 194 genomes were removed as they were very distant in the phylogenetic tree (greater than 80,000 SNPs from K10) and clustered with ATCC13950 or DSM44623 (data not shown).

The final dataset included 1,230 genomes that fell within the quality criteria (**Figure 3**). One exception was made for the single MAS-type strain isolate, which had 808 contigs and was only available from GenBank as an assembly. An assembly metrics summary table is provided within the supplementary tables, **Supplementary Material 4**. The number and broad characteristics of the *M. avium* WGS isolates included in the study are shown in **Table 1**.

## Phylogeny

Whole-genome SNP phylogeny revealed tight clustering of certain subspecies (**Figure 4**). These tight clusters were present in the phylogeny regardless of which reference genome was used to produce the core SNP alignment and tree. The number of isolates in the MAP and MAA/S clades and the individual isolates present were identical between all three trees (supplementary figures, **Supplementary Material 2**). MAP K10 was used for the final tree as it is the most widely used reference genome. A tree produced with the Panaroo MAFFT alignment had similar branch positioning, but branch lengths were more variable within each subspecies.

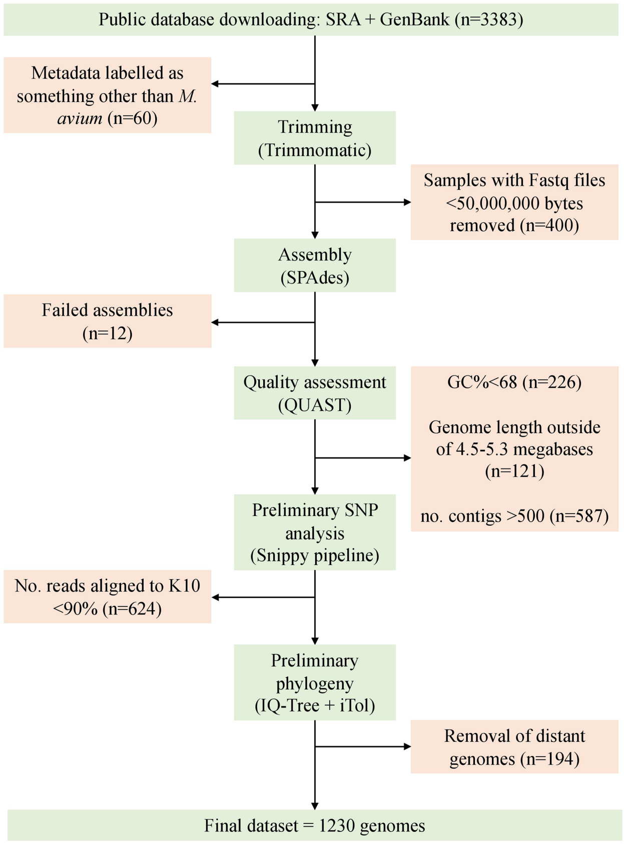

Public database downloading: SRA + GenBank (n=3383)
Metadata labelled as something other than *M. avium* (n=60)
Trimming (Trimmomatic)
Samples with Fastq files <50,000,000 bytes removed (n=400)
Assembly (SPAdes)
Failed assemblies (n=12)
Quality assessment (QUAST)
GC%<68 (n=226)
Genome length outside of 4.5-5.3 megabases (n=121)
no. contigs >500 (n=587)
Preliminary SNP analysis (Snippy pipeline)
No. reads aligned to K10 <90% (n=624)
Preliminary phylogeny (IQ-Tree + iTol)
Removal of distant genomes (n=194)
Final dataset = 1230 genomes

**FIGURE 3** | Overview of the process for sample selection and retention in this study. The final dataset consisted of 1,230 genomes.

MAP genomes formed a distinct clade (green node, **Figure 4**) and together the MAA and MAS subspecies formed another distinct clade (blue node, **Figure 4**). The two known MAS isolates formed a branch within the MAA cluster (see supplementary figures, **Supplementary Material 2** for a more detailed MAA/S phylogeny). MAH was the most diverse of the subspecies in this analysis, with extensive branching and multiple clades present in the phylogenetic tree (**Figure 4**). Two isolates that had been previously typed as MAA, SRR8236370 (Operario et al., 2019) and SRR901356 [also known as *Mycobacterium avium* subsp. *avium* 2,285 (*R*)] are located outside of the MAA/S (MAA and MAS) cluster.

Within the MAP clade, there appeared to be two major lineages and within the smaller of these there appeared to be two sub-lineages: one contained genomes predominantly from the Oceania region, while the other had genomes originating from Europe and America (orange nodes, **Figure 4**). Genomes in the smaller major lineage contained sub-lineages predominantly from sheep and several have been typed as sheep strains of MAP. It is likely that this cluster represents the previously described sheep Type I and Type III sub-lineages of MAP. This is further supported by the presence of the Australian Telford Type I closed sheep genome in one sub-lineage and the S397 (American) and JIII-386 (German) Type III reference strains in the other sub-branch. The larger major lineage within the MAP clade contained isolates that were previously typed as cattle strains (or Type II) and are from predominantly bovine and human hosts. This larger branch within MAP contained the K10 Type II reference genome (see supplementary figures, **Supplementary Material 2** for a more detailed MAP phylogeny).

## SNP Analysis

To identify the presence of SNP hotspots, defined as a 10,000 base-pair region where the number of core SNPs was four SDs greater than the average SNP density across the genome, the location of core genome SNPs was assessed for each subspecies group and then collectively across all *M. avium* genomes (**Table 2**). The reference genome used in the analysis impacted the identification of SNP hotspots. It should be noted that the number of core genome SNPs in the full dataset of 1,230 genomes ("All" category of **Table 2**) did not always equal the number of core genome SNPs detected in the subspecies group analysis, as the core genome differed between the collective *M. avium* group and each of the subspecies groups. This meant that any SNPs specific to the core of one subspecies may not be detected in the full dataset across the different subspecies.

When MAP isolates were mapped to the K10 MAP reference genome, the average core SNP density was 20.6. This was less than half of the average SNP density determined for the other subspecies when mapped to the K10 reference genome, with SNP densities of 97.5 and 189.5 per 10,000 base pairs for MAA/S and MAH, respectively. Similarly, when MAA/S isolates were mapped to the DSM44156 MAA reference genome, the average SNP density per 10,000 base pairs was 15.5, whereas against MAP and MAH isolates, it was 69.1 and 179.1,

**TABLE 1 |** The reported host species, year range, and geographic origin, and the proportions of reported subspecies for 1,230 *Mycobacterium avium* genomes.

| | MAP | MAA/S | MAH | Unreported subspecies | Total |
|---|---|---|---|---|---|
| Africa | 2 | 0 | 0 | 0 | 2 |
| Asia | 10 | 2 | 60 | 0 | 72 |
| America | 400 | 0 | 72 | 106 | 578 |
| Europe | 111 | 3 | 45 | 132 | 291 |
| Oceania | 50 | 0 | 0 | 7 | 57 |
| Not reported | n/a | n/a | n/a | 230 | 230 |
| Total | 523 | 7 | 177 | 475 | 1,230 |
| Host species | Cow (381), camel (2), sheep (70), goat (8), human (8), deer (3), environment (77), Bison (3), and Unknown (16) | Avian (4) and unknown (3) | Avian (2), cow (1), deer (2), environment (23), horse (1), human (149), pig (1), and unknown (3) | *Oryx dammah* (1), avian (46), environment (16), human (181), and unknown (228) | 12 |
| Year range | 1975–2016 | 1901–2015 | 1983–2016 | 1995–2018 | 1901–2018 |
| No. not reported | 43 | 3 | 9 | 233 | 288 |

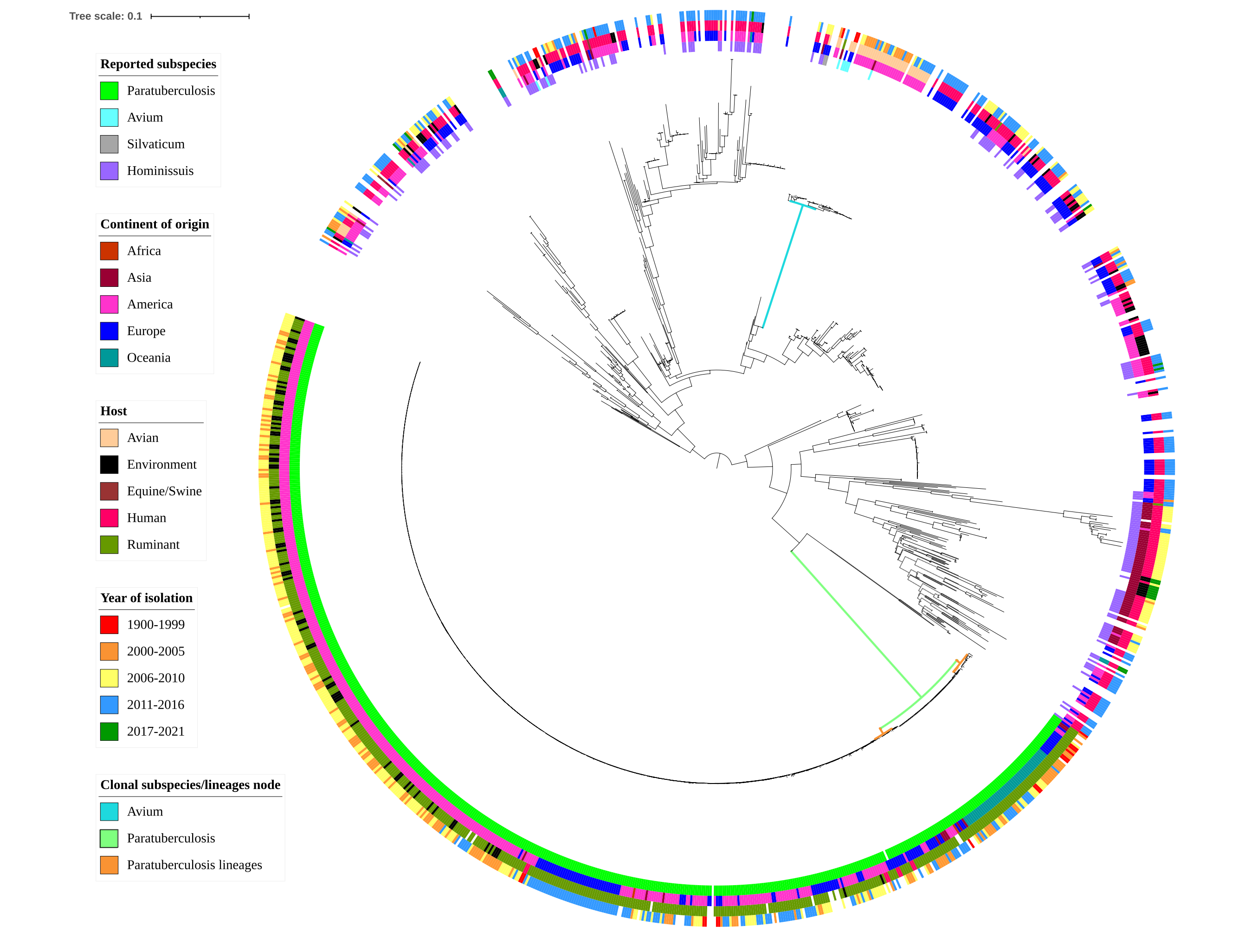


**FIGURE 4 |** Whole-genome core SNP phylogenetic tree and associated metadata of 1,230 *M. avium* isolates based on *M. avium* subspecies *paratuberculosis* (MAP) K10 and rooted to the midpoint. Metadata are depicted by colored circles. Innermost circle is reported subspecies, second circle is the continent of origin of isolate, third circle is host species, and fourth circle is year of isolation. Isolates came from a variety of hosts; thus, some hosts such as avian (such as waterfowl and other poultry) and ruminant (such as sheep and cattle) species were combined within classes. For ease of readability, branch labels were removed. The green node and blue node indicate the MAP and *M. avium* subspecies *avium* (MAA)/S clusters. The smaller orange nodes indicate MAP lineages: the leftmost is the cattle lineage, the middle is the type I sheep lineage and the rightmost is the type III sheep lineage.

respectively. In contrast, when MAH isolates were mapped to the MAH104 reference genome, they had an average SNP density of 156.9. Against the MAH reference genome, MAP and MAA/S isolates had a lower density compared to the

**TABLE 2 |** Number and density of SNPs and SNP hotpots for each subspecies and the full dataset, compared to the K10 (MAP), DSM44156 (MAA), and MAH104 (MAP) reference genomes.

| Subspecies | Reference genome | Total SNPs | Average SNPs per 10,000 bp | Max | SD | No. hotspots | No. of regions with no SNPs |
|---|---|---|---|---|---|---|---|
| MAP | K10 | 88,814 | 20.6 | 49 | 7.5 | 39 | 11 |
| | DSM44156 | 34,286 | 69.1 | 465 | 44.2 | 8 | 26 |
| | MAH104 | 32,850 | 59.9 | 350 | 37.2 | 7 | 77 |
| MAA/S | K10 | 88,814 | 97.5 | 673 | 59.6 | 10 | 16 |
| | DSM44156 | 7,698 | 15.5 | 47 | 6.5 | 22 | 18 |
| | MAH104 | 29,767 | 54.3 | 675 | 56.4 | 6 | 70 |
| MAH | K10 | 91,519 | 189.5 | 655 | 116.1 | 13 | 44 |
| | DSM44156 | 88,814 | 179.1 | 704 | 118.9 | 54 | 12 |
| | MAH104 | 86,006 | 156.9 | 681 | 116.7 | 7 | 102 |
| All* | K10 | 65,313 | 189.5 | 655 | 116.1 | 13 | 44 |
| | DSM44156 | 65,049 | 131.1 | 497 | 92.1 | 9 | 69 |
| | MAH104 | 65,206 | 118.9 | 459 | 92.7 | 8 | 119 |

**Full dataset comprising all 1,230 M. avium isolates.*

MAH subspecies at 59.9 and 54.3 SNPs per 10,000 base pairs, respectively.

## SNP Hotspots

In the full dataset, there were 13, 9, and 8 hotspots (where the number of SNPs present in a 10,000 base-pair region was greater than four standard deviations from the average SNP density) and 44, 69, and 119 regions where no SNPs were present when isolates were mapped to K10 (MAP reference genome), DSM44156 (MAA reference genome), and MAH104 (MAH reference genome), respectively (**Table 2**).

### MAA/S Dataset

SNP analysis demonstrated the presence of several hotspot regions in all subspecies. The most notable were the hotspots in the MAA/S dataset (**Figure 5**). Several major hotspots were seen in these isolates compared to the K10 and MAH104 reference genomes. When the MAA/S dataset was mapped to the K10 reference genome, the largest peaks of 673 and 601 SNPs occurred in reference genome regions 2,410,000–2,420,000 and 2,420,000–2,430,000 base pairs. Within these two 10,000 base-pair regions in K10 are four genes belonging to the mycobactin (mbt) synthesis cluster and several enzymes involved in metabolism. The remainder of the significant peaks occurred at: 1,880,000; 2,240,000; 2,400,000; 3,950,000; 4,140,000; 4,150,000; and 4,240,000. The largest hotspot identified within the MAA/S dataset when compared to the MAH104 reference genome occurred over three 10,000-bp regions between 2,000,000 and 2,030,000 base pairs and contained a total of 1,890 SNPs. This region contained four genes from the mbt cluster, six genes involved in information storage and processing and a further six genes belonging to various Clusters of Orthologous Groups (COG) pathways. Other significant hotspots occurred at 3,530,000 and 5,330,000 bp and contained mainly hypothetical proteins.

### MAP Dataset

Within the MAP dataset, the largest hotspots occurred when compared to the MAA reference genome, DSM44156; these were across two 10,000-bp regions of the reference genome between 2,560,000 and 2,580,000 bp and contained 899 SNPs (**Figure 5**). In these regions of the DSM44156 genome, there are four mbt genes, four genes involved in metabolism, three in transcription and three genes were poorly characterized. Other hotspots in the MAP dataset when compared to the MAA reference genome occurred at: 2,550,000; 4,570,000; 4,580,000; 4,590,000; 4,830,000; and 4,940,000. The largest hotspot for the MAP dataset in comparison with the MAH104 reference genome comprised of 626 SNPs and occurred at base pairs 5,010,000–5,030,000 bp. In this section of the MAH reference genome, there were five genes from various metabolic pathways, four genes involved in transcription and its regulation, two genes involved in cell signaling, and four genes with poor characterization. No other hotspots in the MAP dataset compared to the MAH104 reference exceeded 250 SNPs per 10,000 base pairs; however, some less notable but still significant hotspots occurred at: 2,180,000; 2,380,000; 3,530,000; and 5,460,000.

### MAH Dataset

The MAH dataset had the greatest number of significant peaks in the SNP variant plots and had the highest SNP density compared to all of the reference genomes (**Table 2**; **Figure 5**). When compared to the K10 reference genome, there were 13 significant peak regions ranging from 471 to 655 SNPs per 10,000 base pairs. These peaks were spread throughout the K10 genome and contained genes belonging to a variety of functional categories. The exact locations were at base pairs: 810,000; 870,000; 900,000; 1,060,000; 1,070,000; 3,760,000; 3,980,000; 3,990,000; 4,000,000; 4,010,000; 4,140,000; 4,420,000; and 4,700,000 within K10. For readability, a single asterisk is used where hotspots are located in consecutive 10,000 bp segments ($n = 4$). Ten genes were associated with cellular processes and signaling COGs, 20 were associated with transcription, translation, and replication pathways, 38 genes had unknown functions, and 54 were involved in metabolism. Within the genes involved in metabolic processes, 18 genes were predicted to be involved in lipid transport and metabolism, while 14

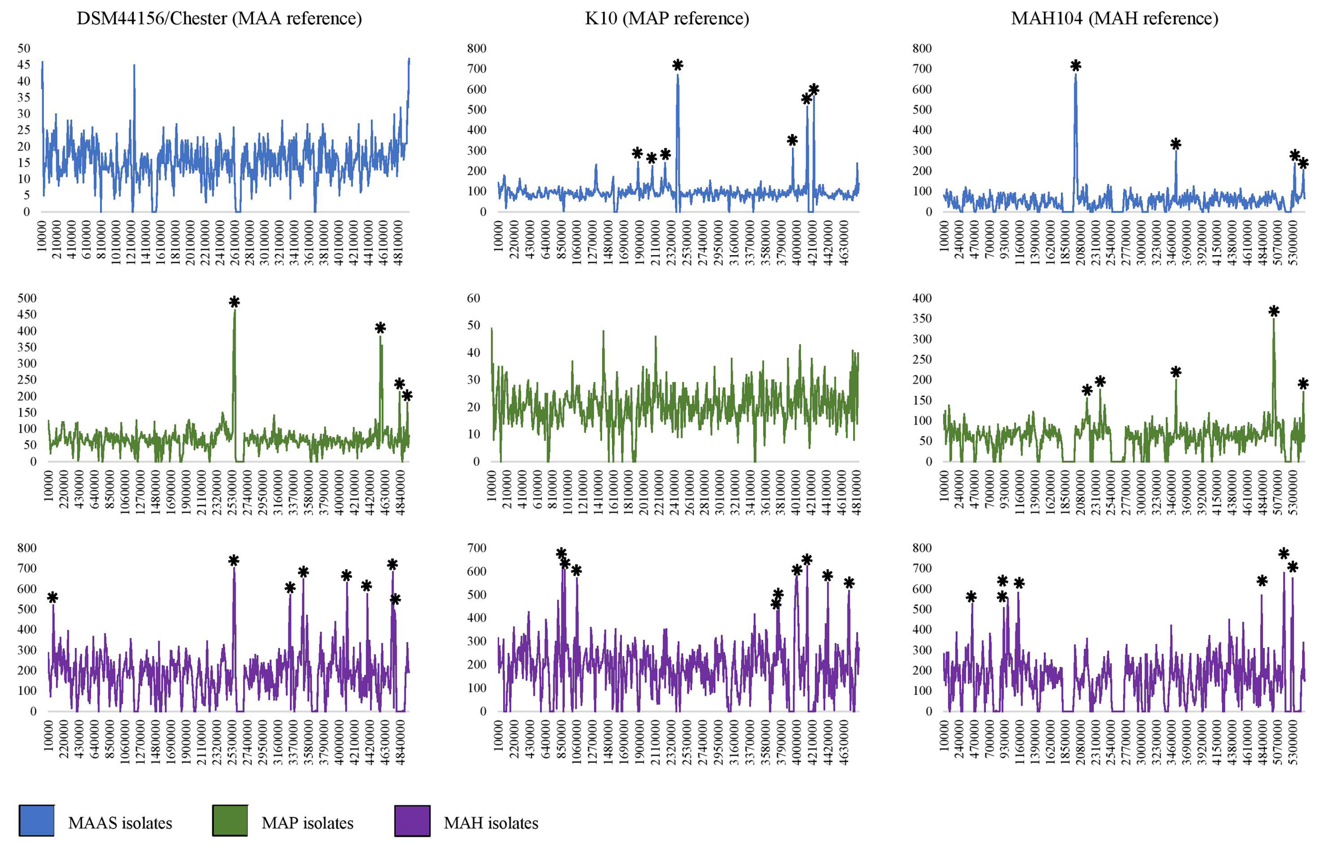


**FIGURE 5 |** Density of SNP variants plots for MAP, MAA/S, *M. avium* subspecies *hominissuis* (MAH), and all isolates compared to closed reference genomes DSM44156 (MAA), K10 (MAP), and MAH104 (MAH). Several regions in each subspecies dataset can be clearly seen to contain several 100 SNPs. Significant peaks are indicated by an asterisk (*). Note where hotspots occurred over consecutive 10,000 bp ranges a single asterisk is used. The *Y* axis indicates the number of SNPs in a 10,000-bp region; the *X* axis is the position in the reference genome starting from the same recognized starting point (*dnaA*).

genes were predicted to have an involvement in secondary metabolite biosynthesis, transport, and catabolism. When compared to the DSM44156 reference genome, the MAH dataset contained 12 significant peaks with 478–704 SNPs per 10,000 bp. These hotspot regions in the MAA reference genome contained 14 genes involved in cellular signaling pathways, 17 involved in information storage and processing, 47 from metabolic pathways, and 33 that were poorly characterized. The largest peak occurred over a region spanning 20,000 bp between 2,560,000 and 2,590,000, with a total of 1,309 SNPs identified. This is the same hotspot region that was identified in the MAP dataset when compared to the MAA reference genome. Other significant peaks occurred at: 80,000; 3,320,000; 3,330,000; 3,510,000; 4,110,000; 4,390,000; 4,730,000; 4,740,000; 4,760,000; and 4,770,000 bp.

*Mycobacterium avium* subspecies *hominissuis* was unique in that it was the only subspecies to exhibit significant peaks when compared to a reference genome that belonged to the same subspecies. Eight significant peaks were found in the MAH dataset when compared to the MAH104 reference genome. These occurred throughout the genome and ranged from 466 to 681 SNPs per 10,000 base pairs. The location in MAH104 with the most SNPs occurred at 5,170,000 base pairs. Other hotspots occurred at: 440,000; 920,000; 990,000; 1,140,000; 4,830,000; and 5,300,000 bp. Prokka annotated nine genes within this region and EggNOG identified two associated with metabolism, three in information storage and processing, one associated with cellular signaling, and two poorly characterized genes.

## Pan-Genome Analysis

*Mycobacterium avium* subspecies *hominissuis* had the smallest number of core genes and the largest number of accessory and cloud genes of all the *M. avium* subspecies (**Figure 6**). There were: 4,055; 182; 1,220; and 8,227 genes in the core, soft core, shell, and cloud and a total number of 13,684 genes. The MAA/S group had a total of 4,688 genes and a breakdown of: 4,174; 220; 140; and 154 genes in the core, soft core, shell, and cloud, respectively. Despite having more than 10 times the number of isolates in the dataset compared to MAA/S and only 31 fewer isolates than MAH, the MAP subspecies had the largest core genome and the smallest number of accessory genes. This indicates a much lower diversity within this subspecies. MAP had a total of 4,777 genes with a core of 4,267 and an accessory genome 75, 72, and 363 genes in the soft core, shell, and cloud. Overall, the proportion of genes in each category was similar in MAA/S and MAP, where the core genome made up 89% of the total number of genes. In contrast, MAH had only 30% of genes that were classified as being in the core and the majority of genes (60%) were within the cloud genome (**Figure 6** and supplementary tables, **Supplementary Material 4**).

## Subspecies-Specific Marker Genes

Scoary software identified 76, 37 and three subspecies-specific markers for MAP, MAA/S, and MAH, respectively. Of these, 46, 27, and two genes for MAP, MAA/S, and MAH reached the BLAST thresholds of a matching identity greater than or equal to 97%, *e*-value less than 0.000001, bitscore of 100, or greater and coverage of 95% of the full gene length or higher. EggNOG software demonstrated that these marker genes came from a variety of functional categories. Of the genes that had been validated with BLAST, 37 and 13 genes from MAP and MAA/S and the two from MAH were successfully functionally annotated by EggNOG (**Figure 7**, supplementary data, **Supplementary Material 3**).

In MAP, most of the subspecies-specific marker genes belonged to the metabolism COG category. Specifically, 15 MAP-specific genes were on COG pathways that are involved in inorganic ion and secondary metabolite biosynthesis, transport, and catabolism (categories P and Q; Galperin et al., 2021). The majority of the markers found in MAA/S were of unknown function (19 of 27); the genes that were known were associated with metabolism or information storage and processing (**Figure 7**). Both of the two species-specific markers in MAH were involved in metabolic pathways. One of the genes, a fumarylacetoacetate (FAA) hydrolase, belonged specifically to category Q; secondary metabolites biosynthesis, transport, and catabolism pathways. The other gene was a glyoxalase bleomycin resistance protein dioxygenase (bphC_1) and belonged to the amino acid transport and metabolism (category E).

# DISCUSSION

This investigation aimed to describe genetic differences between the *M. avium* subspecies using a pan-genome and mapping approach. The dataset included 1,230 genomes from a variety of geographical locations, host species, and years. This included 21 isolates sequenced specifically for this investigation and deposited in the SRA. To our knowledge, this dataset represents the largest number of *M. avium* genomes that has been used

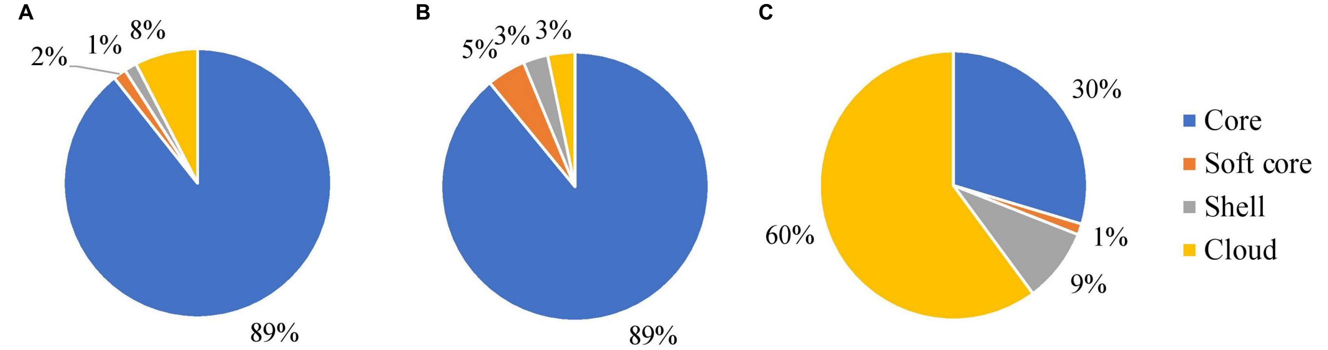


**FIGURE 6 |** The relative proportions of genes belonging to the core, soft core, shell, and cloud genomes of MAP **(A)**, MAA/S **(B)**, and MAH **(C)** datasets. A table containing the number of genes in each section is available in the supplementary tables, **Supplementary Material 4**.

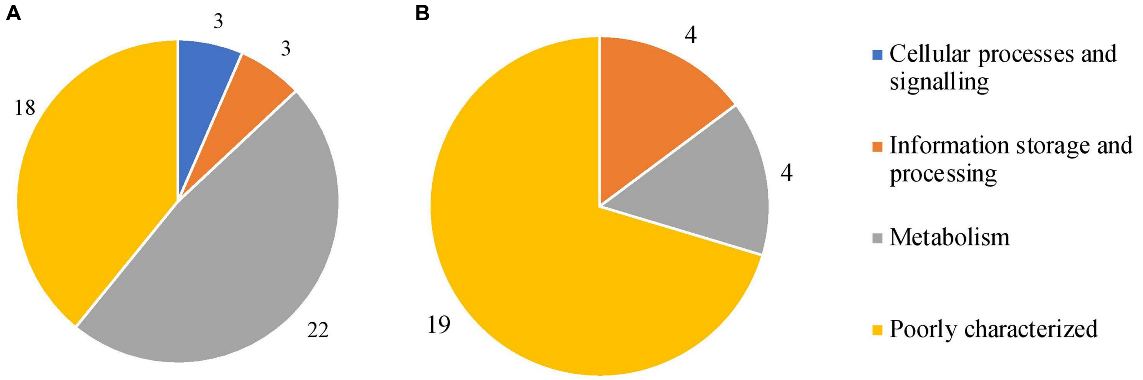


**FIGURE 7 |** Clusters of Orthologous Groups (COG) categories for all marker genes in MAP **(A)** and MAA/S **(B)**. These genes were present in all whole-genome sequencing (WGS) isolates from their respective subspecies (MAP $n = 575$, MAA/S $n = 49$) and absent from all other subspecies.

for comparative genomics of these subspecies. Furthermore, the large phylogeny that this study provides may be useful for outbreak tracing, should an *M. avium* outbreak occur. The novel SNP hotspot approach that was undertaken enabled the identification of regions that have undergone extensive mutation during niche adaptation. Subspecies-specific putative virulence genes based on homologies to genes with similar functions to other taxa were identified by the pan-genome analysis. The core SNP phylogeny revealed tight clustering of MAP and MAA/S subspecies and high diversity of MAH. The phenomena of niche adaptation by these subspecies appears to have occurred by a combination of mutations within regions common to all subspecies and the loss or gain of additional genes required for survival in their respective environments.

Phylogenomic analysis demonstrated a clear distinction between subspecies. Many of these isolates had been described only to the species level and had not been subtyped, but all known MAA/S and MAP subspecies clustered together. MAA and MAS formed one tight clade while MAP formed another, separate, tight clade. Two isolates that had been previously typed as MAA, SRR8236370 (Operario et al., 2019) and SRR901356 [also known as *Mycobacterium avium* subsp. *avium* 2,285 (*R*)] were located outside of the MAA/S (MAA and MAS) cluster, suggesting that they had been misclassified originally. This is supported by previous investigation where 2285R was seen within a clade of MAH genomes (Uchiya et al., 2013). MAH demonstrated high genomic diversity and did not form a single clade. Known MAH isolates were scattered throughout the phylogeny but were absent from the MAP and MAA/S branches. An earlier study suggested that MAA/MAS and MAP are pathogenic clones within *M. avium* (Turenne et al., 2008). This is supported by the phylogeny of the present study, where the MAP and MAA/S appeared to be sub-clades within MAH. MAP is considered the most clonal of the subspecies with MAA intermediate but much less diverse than MAH (Kei-Ichi et al., 2017). Other investigators also reported clustering of subspecies (Kei-Ichi et al., 2017; Bannantine et al., 2020a). When MAP and non-MAP genomes underwent phylogenomic analysis, MAA strains were intermediate to MAH and MAP strains and two distinct clusters, one MAP and one non-MAP were present (Bannantine et al., 2020a). A slightly larger dataset demonstrated three distinct clusters, one consisting of MAH isolates, another of MAA and the MAS-type strain and a final cluster split into two sub-lineages with MAP isolates contained within one sub-lineage and two MAH isolates (A5 and 10-4249) in the remaining sub-lineage (Kei-Ichi et al., 2017). The two MAH isolates (A5 and 10-4249) that were close to MAP in Kei-Ichi and colleagues' investigation were also situated close to MAP isolates in the present study along with several other MAH genomes.

The midpoint root in the phylogeny resulted in two distinct branches of *M. avium*. One contained roughly half of the MAH isolates and the MAA/S cluster and the other contained the remaining MAH isolates and the MAP cluster. Suggestions to subdivide MAH into multiple subspecies have been put forward (Turenne et al., 2008), and the results of the present study support this proposal. Identification of specific lineages associated with virulent genotypes or host tropism provides an opportunity to understand how pathogens evolve. Genomic differences between lineages of MAP have been studied extensively. Numerous investigations implicate lineage-specific regions and mutations as a reason for phenotypic differences between strains (Bannantine et al., 2012; Wibberg et al., 2020; Lim et al., 2021). Knowledge of strain-specific characteristics that are linked with virulence or host tropism is crucial for understanding pathogenesis. Applying similar methods to lineages of MAH within the current dataset may provide valuable insights into how MAH is able to cause disease in some instances. Further work to investigate differences between MAA-like and MAP-like MAH isolates separated by the midpoint may reveal novel insights on pathogenicity and host-specificity for these groups and provide phenotypic support for the split of MAH. This may reveal gene or SNP associations with certain phenotypes or lineages. Such investigations on a smaller scale have been undertaken. Several loci specific to the hypervirulent TH135 MAH reference genome were associated with isolates from patients with progressive pulmonary disease (Kei-Ichi et al., 2017). The authors suggest these distinct features may have been acquired during strain evolution and could play an important role in the progression of clinical MAC disease. Distinct genetic features have been found between isolates that display increased virulence or progressive disease in immunocompetent patients (Uchiya et al., 2013) including plasmids (Takahashi et al., 2015) and specific lineages (Kei-Ichi et al., 2017). Further *in vitro* work may assist in improving our understanding of virulent genotypes in various regions and countries. Knowledge in this area may improve patient prognosis and treatment outcomes.

The SNP and pan-genome analyses told a similar story, with MAH considered the most diverse subspecies and MAA/S and MAP considerably less diverse. The largest accessory genome and smallest core genome (30%) were found in MAH isolates. In contrast, pathogenic subspecies MAP and MAA/S have a large core (89%) and comparatively smaller accessory component. Similar pan-genome findings were reported previously for MAP and MAH (Bannantine et al., 2020a,b). The inherent diversity between MAH isolates was also evident in the mutation hotspot analysis, with multiple SNP hotspots identified when the MAH dataset was compared to a MAH reference genome. In contrast, no significant hotspots were seen in MAP and MAA/S datasets when mapped to a reference genome of their own subspecies, indicative of a higher clonality within these subspecies. Further, MAH had a high number of SNPs compared to all reference genomes. The high level of diversity within the MAH subspecies and the very limited diversity of MAP, MAA, and MAS has been demonstrated in other studies (Turenne et al., 2008; Pate et al., 2011; Kei-Ichi et al., 2017; Bannantine et al., 2020a) and may reflect the saprophytic lifestyle of MAH; it is commonly found in the environment and likely to be subject to a wide range of conditions. In contrast, obligate pathogens, such as MAA, MAS, and MAP are exposed to a narrow range of conditions due to an intracellular lifestyle and are adapted to that environment. Thus, a high level of genomic diversity may not be necessary for survival of MAA, MAS, and MAP.

The present study found many subspecies-specific coding sequences in the MAA/S and MAP subspecies that may play a role in niche adaptation. Comparisons between the pathogenic subspecies (MAA, MAS, and MAP) and MAH revealed no common genes present in the clonal, pathogenic species that were absent in MAH. This may reflect the phenotype of gut infection in MAP vs. lung infection in MAA/S requiring different sets of virulence genes for successful infection in these tissues. Other studies have looked at subspecies-specific loci, and several genes have been found that match between studies. A group in Japan utilized 79 *M. avium* genomes and found several subspecies-specific loci containing virulence genes (Kei-Ichi et al., 2017). MAP and MAA/S were missing one locus that contains PPE proteins. A MAP-specific locus containing several coding sequences for genes encoding Mec, MmpL/MmpS, and PPE proteins was also discovered (Kei-Ichi et al., 2017). MAH-specific PPE proteins and MAP-specific Mec, MmpL/MmpS were not found in the present study; however, a MAP-specific PPE protein was identified. More recently, Bannantine and colleagues used a pan-genome approach on 29 closed *M. avium* genomes. They reported 86 genes specific to MAP, seven specific to MAA, and three that were specific to MAH (Bannantine et al., 2020b). Subspecies-specific loci and genes are likely to be associated with adaptation to human and porcine hosts by MAH or ruminant hosts by MAP.

Subspecies-specific markers found in the MAA/S subspecies were predominantly hypothetical proteins. The coding sequences that were successfully annotated were involved in metabolism, specifically energy production/conversion and transcription. Metabolic pathways were also represented in the SNPs analysis, with hotspots in the MAA/S dataset consistently in regions of the reference genomes that contained genes encoding enzymes involved in metabolism and mycobactin (mbt) synthesis. Previous work on metabolic pathways in MAA has demonstrated a requirement of cholesterol for virulence (De Chastellier and Thilo, 2006). The utilization of cholesterol and fatty acids has also been described in MAP (Weigoldt et al., 2013; Thirunavukkarasu et al., 2014), but no direct comparison of these two subspecies has been undertaken. Findings from a smaller, closed genome dataset revealed seven genes specific to MAA; several of these genes overlap with MAA-specific genes found in the present study. Future studies to uncover the function of the MAA/S-specific genes may reveal unique pathways that confer a survival advantage in avian hosts.

Within the SNP analysis, several genes from the mbt cluster were found in hotspot regions. In the MAA/S dataset, hotspots were found to the K10 (MAP) and MAH104 (MAH) reference genomes within regions that contained mbt cluster genes. The same mbt cluster genes were also found in hotspots within the MAP and MAH datasets when they were compared to the DSM44156 MAA reference genome. Studies in *M. avium* and other mycobacteria have demonstrated the importance of mycobactin and iron utilizing pathways for survival and virulence (De Voss et al., 1999; Fang et al., 2015). MAA and MAS may have adaptations in iron sequestering pathways that offer a survival advantage in their preferred avian hosts. Further *in vitro* experiments may be needed to fully appreciate the consequences of mutations present in these genes.

The largest subcategory of MAP-specific genes in the present study was associated with secondary metabolites biosynthesis, transport, and catabolism. A similar finding was also reported by another recent investigation (Bannantine et al., 2020b). Several genes were also annotated as mammalian cell entry (Mce) genes. Mce genes are involved in invasion and persistence within host cells (Chitale et al., 2001; El-Shazly et al., 2007; Chandolia et al., 2014). Interactions with host cell receptors have been described (Zhang et al., 2018) and may reflect the unique host and tissue tropisms of this subspecies. MAP has been the most extensively studied in the search for subspecies-specific genes (Gold et al., 2001; Rodriguez et al., 2002; Li et al., 2005; Janagama et al., 2009; Wang et al., 2014, 2016). Earlier investigations found some of the MAP-specific genes that were identified in the present study including a cytochrome P450, polyketide synthase, and several other genes predominantly involved in iron acquisition and metabolism. MAP is mycobactin dependant in culture conditions and has evolved MAP-specific iron sequestering pathways (Barclay et al., 1985; Collins et al., 1985). No MAP-specific genes that were annotated as mbt genes were identified in the present study. However, the large number of genes involved in secondary metabolites biosynthesis indicates there may be MAP-specific genes involved in this pathway. Further work interrogating these MAP-specific genes *in vitro* is required to reveal their functions to see if they are indeed linked to iron-sequestering pathways. Furthermore, these genes may be essential for the unique gut tropism of this subspecies, and additional study may reveal its virulence mechanisms.

Genes or SNPs that are specific to a particular subspecies may be suitable for use as diagnostic markers. Current methods used to differentiate *M. avium* subspecies are typically based on RFLP patterns. Various insertion sequences (IS) are exploited for this purpose including IS*1245* and IS*1311* for MAH (Mijs et al., 2002), IS*901* in MAA and MAS (absent from MAH) or the MAP-specific IS*900* (Moravkova et al., 2008; Rindi and Garzelli, 2014). SNPs within these IS elements can also be used to differentiate lineages within certain subspecies, such as IS*1311* in MAP (Whittington et al., 1998). IS*1311* shares 85% sequence identity with IS*1245* (Johansen et al., 2005), and IS*900* has apparent homologues in other mycobacterial species (Cousins et al., 1999). Variable number tandem repeat (VNTRs) typing using MIRUs has been developed and has higher resolution than PCRs based on insertion elements (Radomski et al., 2010; Rindi and Garzelli, 2014). However, sometimes this technique is insufficient for closely related isolates and a combination of techniques is required to distinguish between bird-type isolates (Pate et al., 2011). Use of whole genes for diagnostic tests may be more accurate than using insertion sequences or SNPs in particular genes. The recent publication by Bannantine and colleagues (Bannantine et al., 2020b) and the complementary results from the present investigation demonstrate there are a variety of subspecies-specific genes that could be used diagnostic targets. Further work would be required to determine the suitability of these marker genes in a larger dataset as candidates for diagnostic targets.

The main limitations of the study include use of *de novo* assemblies, culling of genomes that did not meet assembly standards (~30% of assemblies), strict cutoffs to call marker genes and a small number of MAA and MAS genomes. These limitations may have resulted in reduced apparent genomic diversity within the MAA and MAS genomes and fewer features relevant to MAA and MAS subspecies. For example, the inclusion of more MAS genomes may have revealed another distinct cluster within the MAA subspecies. Further, an absence of good coverage across some areas of genomes due to *de novo* assembly and strict criteria for marker genes to be called means that other subspecies-specific genes and loci likely exist. However, the thresholds that were used in this study minimized the chance of a gene marker being the result of an assembly error and improved the validity of the pan-genome results. Mycobacterial isolates are notoriously difficult to culture due to a slow growth rated compared to other bacterial species. Their sequencing and assembly is also difficult due to the presence of repetitive regions of high GC content. To overcome assembly and annotation artifacts, other researchers limited their study to closed genomes only (Bannantine et al., 2020a), an approach that would accurately demonstrate that the larger number of genes in the accessory genome of non-MAP subspecies is not due to assembly artifact but rather is a true reflection of the differences in diversity between MAH and other *M. avium* subspecies, which validates the pan-genome findings of the present study.

The use of a reference genome that belonged to a single subspecies for use in the final analysis may have introduced some bias within the phylogeny. Future *M. avium* studies should consider the use of a bioinformatically created most recent common ancestor (MRCA) as has been undertaken for the Mycobacterium tuberculosis complex (Comas et al., 2013). This method may reveal insights in the historic spread of *M. avium* globally and aid in controlling future transmission.

## CONCLUSION

Increasing our understanding of the genome of *M. avium* subspecies will lead to insights into mycobacterial virulence, pathogen evolution, host preference, and tissue tropisms in pathogenesis. This new information is vital for understanding the clinical significance of the MAC for human and animal health and will lead to improvements in diagnosis, control, and treatment. We confirm earlier findings from Turenne et al. (2008) that *Mycobacterium avium* is a species made up of one highly diverse subspecies, known as MAH, and at least two pathogenic sub-clones, namely MAA and MAP, that have adapted to specific host niches. Due to the small number of MAS isolates available, no conclusions could be drawn for this subspecies except that it is closely related to MAA. The mapping approach revealed several areas in each subspecies where extensive mutations have occurred relative to a reference genome from other subspecies. Hotspots occurred in regions where known mycobacterial virulence genes were present in the reference genome. The pan-genome analysis confirmed that MAH is highly diverse, whereas MAA/S and MAP are quite clonal, with the MAP subspecies being the most clonal. Several subspecies-specific coding sequences were found that belong to a variety of COG categories. These differences between subspecies may reflect their adaptation to different lifestyles.

## DATA AVAILABILITY STATEMENT

The whole genome data generatated for this study can be found in the Sequence Read Archive https://www.ncbi.nlm.nih.gov/sra under BioProject ID PRJNA809746.

## AUTHOR CONTRIBUTIONS

RM was responsible for the study design, dataset curation, bioinformatics analysis, and writing of the manuscript. VT and KP contributed to the study design and editing of the manuscript. RW assisted in the curation and description of genomes sequenced by the University of Sydney and editing of the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by Meat and Livestock Australia (MLA; grant number P.PSH.0813). The funding body (MLA) provided support in the form of some authors' salaries and research materials but did not have any role in the study design, data collection, analysis, or preparation of the manuscript.

## ACKNOWLEDGMENTS

The authors would like to acknowledge laboratory assistance from Anna Waldron and bioinformatic assistance from Keenan Pey and Winkie Fong. David Dawson, Senior Scientist, Mycobacterium Reference Laboratory, Queensland Health Pathology Services kindly provided 20 *M. avium* isolates to RW in 2001 for ongoing research. The authors would also like to acknowledge the Sydney Informatics Hub for enabling the use of the Artemis supercomputer cluster and for training received by the RM.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.892333/full#supplementary-material

## REFERENCES

Bach, H. (2015). What role does *mycobacterium avium* subsp. *Paratuberculosis* play in crohn's disease? *Curr. Infect. Dis. Rep.* 17:463. doi: 10.1007/s11908-015-0463-z

Bang, D., Herlin, T., Stegger, M., Andersen, A. B., Torkko, P., Tortoli, E., et al. (2008). *Mycobacterium arosiense* sp. Nov., a slowly growing, scotochromogenic species causing osteomyelitis in an immunocompromised child. *Int. J. Syst. Evol. Microbiol.* 58, 2398–2402. doi: 10.1099/ijs.0.65503-0

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

Bannantine, J. P., Conde, C., Bayles, D. O., Branger, M., and Biet, F. (2020a). Genetic diversity among *Mycobacterium avium* subspecies revealed by analysis of complete genome sequences. *Front. Microbiol.* 11:1701. doi: 10.3389/fmicb.2020.01701

Bannantine, J. P., Stabel, J. R., Bayles, D. O., Conde, C., and Biet, F. (2020b). Diagnostic sequences That Distinguish M. avium subspecies strains. *Front. Vet. Sci.* 7:620094. doi: 10.3389/fvets.2020.620094

Bannantine, J. P., Wu, C.-w., Hsu, C., Zhou, S., Schwartz, D. C., Bayles, D. O., et al. (2012). Genome sequencing of ovine isolates of *mycobacterium avium* subspecies *paratuberculosis* offers insights into host association. *BMC Genomics* 13:89. doi: 10.1186/1471-2164-13-89

Barclay, R., Ewing, D. F., and Ratledge, C. (1985). Isolation, identification, and structural analysis of the mycobactins of *mycobacterium avium*, *mycobacterium intracellulare*, *mycobacterium scrofulaceum*, and *mycobacterium paratuberculosis*. *J. Bacteriol.* 164, 896–903. doi: 10.1128/jb.164.2.896-903.1985

Ben Salah, I., Cayrou, C., Raoult, D., and Drancourt, M. (2009). *Mycobacterium marseillense* sp. Nov., *mycobacterium timonense* sp. Nov. And *mycobacterium bouchedurhonense* sp. Nov., members of the mycobacterium avium complex. *Int. J. Syst. Evol. Microbiol.* 59, 2803–2808. doi: 10.1099/ijs.0.010637-0

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

Brown-Elliott, B. A., Nash, K. A., and Wallace, R. J. (2012). Antimicrobial susceptibility testing, drug resistance mechanisms, and therapy of infections with nontuberculous mycobacteria. *Clin. Microbiol. Rev.* 25, 545–582. doi: 10.1128/CMR.05030-11

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

Chandolia, A., Rathor, N., Sharma, M., Saini, N. K., Sinha, R., Malhotra, P., et al. (2014). Functional analysis of mce4a gene of *mycobacterium tuberculosis* h37rv using antisense approach. *Cell. Microbiol.* 169, 780–787. doi: 10.1016/j.micres.2013.12.008

Chitale, S., Ehrt, S., Kawamura, I., Fujimura, T., Shimono, N., Anand, N., et al. (2001). Recombinant *mycobacterium tuberculosis* protein associated with mammalian cell entry. *Cell. Microbiol.* 3, 247–254. doi: 10.1046/j.1462-5822.2001.00110.x

Collins, P., McDiarmid, A., Thomas, L. H., and Matthews, P. R. J. (1985). Comparison of the pathogenicity of *mycobacterium paratuberculosis* and mycobacterium spp isolated from the wood pigeon (*columba palumbus-l*). *J. Comp. Pathol.* 95, 591–597. doi: 10.1016/0021-9975(85)90028-3

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., et al. (2013). Out-of-africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* 45, 1176–1182. doi: 10.1038/ng.2744

Cousins, D. V., Whittington, R., Marsh, I., Masters, A., Evans, R. J., and Kluver, P. (1999). Mycobacteria distenct from *mycobacterium avium* subsp. *Paratuberculosis* isolated from the faeces of ruminants possess is900-like sequences detectable is900 polymerase chain reaction: implications for diagnosis. *Mol. Cell. Probes* 13, 431–442. doi: 10.1006/mcpr.1999.0275

De Chastellier, C., and Thilo, L. (2006). Cholesterol depletion in *mycobacterium avium* -infected macrophages overcomes the block in phagosome maturation and leads to the reversible sequestration of viable mycobacteria in phagolysosome-derived autophagic vacuoles. *Cell. Microbiol.* 8, 242–256. doi: 10.1111/j.1462-5822.2005.00617.x

De Voss, J. J., Rutter, K., Schroeder, B. G., and Barry, C. E. III (1999). Iron acquisition and metabolism by mycobacteria. *J. Bacteriol.* 181, 4443–4451. doi: 10.1128/JB.181.15.4443-4451.1999

Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J.-P., et al. (2011). Sniplay: A web-based tool for detection, management and analysis of snps. Application to grapevine diversity projects. *BMC Bioinformatics* 12:134. doi: 10.1186/1471-2105-12-134

Didelot, X., Pollock, S., Tang, P., Crisan, A., Johnston, J., Colijn, C., et al. (2016). Declaring a tuberculosis outbreak over with genomic epidemiology. *Microb. Genom.* 2:e000060. doi: 10.1099/mgen.0.000060

Ekundayo, T. C., Falade, A. O., Igere, B. E., Iwu, C. D., Adewoyin, M. A., Olasehinde, T. A., et al. (2022). Systematic and meta-analysis of *mycobacterium avium* subsp. *Paratuberculosis* related type 1 and type 2 diabetes mellitus. *Sci. Rep.* 12:4608. doi: 10.1038/s41598-022-08700-4

El-Shazly, S., Ahmad, S., Mustafa, A., Al-Attiyah, R., and Krajci, D. (2007). Internalization by hela cells of latex beads coated with mammalian cell entry (mce) proteins encoded by the mce3 operon of *mycobacterium tuberculosis*. *J. Med. Microbiol.* 56, 1145–1151. doi: 10.1099/jmm.0.47095-0

Eslami, M., Shafiei, M., Ghasemian, A., Valizadeh, S., Al-Marzoqi, A. H., Shokouhi Mostafavi, S. K., et al. (2019). *Mycobacterium avium* paratuberculosis and mycobacterium avium complex and related subspecies as causative agents of zoonotic and occupational diseases. *J. Cell. Physiol.* 234, 12415–12421. doi: 10.1002/jcp.28076

Fang, Z., Sampson, S. L., Warren, R. M., Gey van Pittius, N. C., and Newton-Foot, M. (2015). Iron acquisition strategies in mycobacteria. *Tuberculosis* 95, 123–130. doi: 10.1016/j.tube.2015.01.004

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

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* 364, 730–739. doi: 10.1056/NEJMoa1003176

Goethe, R., Laarmann, K., Spröer, C., and Bunk, B. (2020). Complete genome sequence of Mycobacterium avium subsp.aviumChester (DSM 44156). *Microbiol. Res. Announc.* 9, e01549–e01619. doi: 10.1128/MRA.01549-19

Gold, B., Rodriguez, G. M., Marras, S. A. E., Pentecost, M., and Smith, I. (2001). The *mycobacterium tuberculosis* Ider is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. *Mol. Microbiol.* 42, 851–865. doi: 10.1046/j.1365-2958.2001.02684.x

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

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

Janagama, H. K., Senthilkumar, T. M. A., Bannantine, J. P., Rodriguez, G. M., Smith, I., Paustian, M. L., et al. (2009). Identification and functional characterization of the iron-dependent regulator (Ider) of *mycobacterium avium* subsp. *Paratuberculosis*. *Microbiology* 155, 3683–3690. doi: 10.1099/mic.0.031948-0

Johansen, T. B., Djonne, B., Jensen, M. R., and Olsen, I. (2005). Distribution of is1311 and is1245 in *mycobacterium avium* subspecies revisited. *J. Clin. Microbiol.* 43, 2500–2502. doi: 10.1128/JCM.43.5.2500-2502.2005

Kei-Ichi, U., Shuta, T., Taku, N., Shoki, A., Toshiaki, N., and Kenji, O. (2017). Comparative genome analyses of *mycobacterium avium* reveal genomic features of its subspecies and strains that cause progression of pulmonary disease. *Sci. Rep.* 7:39750. doi: 10.1038/srep39750

Kim, B.-J., Choi, B.-S., Lim, J.-S., Choi, I.-Y., Lee, J.-H., Chun, J., et al. (2012). Complete genome sequence of *mycobacterium intracellulare* strain atcc 13950t. *J. Bacteriol.* 194:2750. doi: 10.1128/JB.00295-12

Kim, B.-J., Math, R. K., Jeon, C. O., Yu, H.-K., Park, Y.-G., Kook, Y.-H., et al. (2013). *Mycobacterium yongonense* sp. Nov., a slow-growing non-chromogenic species closely related to *mycobacterium intracellulare*. *Int. J. Syst. Evol. Microbiol.* 63, 192–199. doi: 10.1099/ijs.0.037465-0

Lee, S.-Y., Kim, B.-J., Kim, H., Won, Y.-S., Jeon, C. O., Jeong, J., et al. (2016a). *Mycobacterium paraintracellulare* sp. Nov., for the genotype int-1 of *mycobacterium intracellulare*. *Int. J. Syst. Evol. Microbiol.* 66, 3132–3141. doi: 10.1099/ijsem.0.001158

Lee, I., Ouk Kim, Y., Park, S.-C., and Chun, J. (2016b). Orthoani: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi: 10.1099/ijsem.0.000760

Letunic, I., and Bork, P. (2019). Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Li, L., Bannantine, J. P., Zhang, Q., Amonsin, A., May, B. J., Alt, D., et al. (2005). The complete genome sequence of *mycobacterium avium* subspecies paratuberculosis. *Proc. Nat. Acad. Sci. U.S.A.* 102, 12344–12349. doi: 10.1073/pnas.0505662102

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

Lim, J., Park, H.-T., Ko, S., Park, H.-E., Lee, G., Kim, S., et al. (2021). Genomic diversity of mycobacterium avium subsp. *Paratuberculosis*: pangenomic approach for highlighting unique genomic features with newly constructed complete genomes. *Vet. Res.* 52:46. doi: 10.1186/s13567-021-00905-1

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Goker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14:60. doi: 10.1186/1471-2105-14-60

Mijs, W., De Haas, P., Rossau, R., Van Der Laan, T., Rigouts, L., Portaels, F., et al. (2002). Molecular evidence to support a proposal to reserve the designation *mycobacterium avium* subsp. *Avium* for bird-type isolates and *'m. Avium* subsp. *Hominissuis'* for the human/porcine type of *m. Avium*. *Int. J. Syst. Evol. Microbiol.* 52, 1505–1518. doi: 10.1099/00207713-52-5-1505

Mizzi, R., Timms, V. J., Price-Carter, M. L., Gautam, M., Whittington, R., Heuer, C., et al. (2021). Comparative genomics of Mycobacterium avium subspecies Paratuberculosis sheep strains. *Front. Vet. Sci.* 8:637637. doi: 10.3389/fvets.2021.637637

Möbius, P., Hölzer, M., Felder, M., Nordsiek, G., Groth, M., Köhler, H., et al. (2015). Comprehensive insights in the *mycobacterium avium* subsp. *Paratuberculosis* genome using new wgs data of sheep strain jiii-386 from Germany. *Genome Biol. Evol.* 7, 2585–2601. doi: 10.1093/gbe/evv154

Moravkova, M., Hlozek, P., Beran, V., Pavlik, I., Preziuso, S., Cuteri, V., et al. (2008). Strategy for the detection and differentiation of *mycobacterium avium* species in isolates and heavily infected tissues. *Res. Vet. Sci.* 85, 257–264. doi: 10.1016/j.rvsc.2007.10.006

Moravkova, M., Lamka, J., Slany, M., and Pavlik, I. (2013). Genetic is 901 rflp diversity among *mycobacterium avium* subsp. *Avium* isolates from four pheasant flocks. *J. Vet. Sci.* 14, 99–102. doi: 10.4142/jvs.2013.14.1.99

Murcia, M., Tortoli, E., Menendez, M., Palenque, E., and Garcia, M. (2006). *Mycobacterium colombiense* sp. Nov., a novel member of the mycobacterium avium complex and description of mac-x as a new its genetic variant. *Int. J. Syst. Evol. Microbiol.* 56, 2049–2054. doi: 10.1099/ijs.0.64190-0

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

Operario, D. J., Pholwat, S., Koeppel, A. F., Prorock, A., Bao, Y., Sol-Church, K., et al. (2019). Mycobacterium avium complex diversity within lung disease, as revealed by whole-genome sequencing. *Am. J. Respir. Crit. Care Med.* 200, 393–396. doi: 10.1164/rccm.201903-0669LE

Pate, M., Kušar, D., Žolnir-Dovč, M., and Ocepek, M. (2011). Miru–vntr typing of *mycobacterium avium* in animals and humans: heterogeneity of *mycobacterium avium* subsp. *Hominissuis* versus homogeneity of *mycobacterium avium* subsp. *Avium* strains. *Res. Vet. Sci.* 91, 376–381. doi: 10.1016/j.rvsc.2010.10.001

Paustian, M., Kapur, V., and Bannantine, J. P. (2005). Comparative genomic hybridizations reveal genetic regions within the mycobacterium avium complex that are divergent from *mycobacterium avium* subsp. *Paratuberculosis* isolates. *J. Bacteriol.* 187, 2406–2415. doi: 10.1128/JB.187.7.2406-2415.2005

Paustian, M., Zhu, X., Sreevatsan, S., Robbe-Austerman, S., Kapur, V., and Bannantine, J. P. (2008). Comparative genomic analysis of *mycobacterium avium* subspecies obtained from multiple host species. *BMC Genomics* 9:135. doi: 10.1186/1471-2164-9-135

Radomski, N., Thibault, V. C., Karoui, C., de Cruz, K., Cochard, T., Gutiérrez, C., et al. (2010). Determination of genotypic diversity of *mycobacterium avium* subspecies from human and animal origins by mycobacterial interspersed repetitive-unit-variable-number tandem-repeat and is1311 restriction fragment length polymorphism typing methods. *J. Clin. Microbiol.* 48, 1026–1034. doi: 10.1128/JCM.01869-09

Rindi, L., and Garzelli, C. (2014). Genetic diversity and phylogeny of *mycobacterium avium*. *Infect. Genet. Evol.* 21, 375–383. doi: 10.1016/j.meegid.2013.12.007

Riojas, M. A., McGough, K. J., Rider-Riojas, C. J., Rastogi, N., and Hazbón, M. H. (2018). Phylogenomic analysis of the species of the mycobacterium tuberculosis complex demonstrates that *mycobacterium africanum, mycobacterium bovis, mycobacterium caprae, mycobacterium microti and mycobacterium pinnipedii* are later heterotypic synonyms of *mycobacterium tuberculosis*. *Int. J. Syst. Evol. Microbiol.* 68, 324–332. doi: 10.1099/ijsem.0.002507

Rodriguez, G. M., Voskuil, M. I., Gold, B., Schoolnik, G. K., and Smith, I. (2002). Ider, an essential gene in *mycobacterium tuberculosis*: role of Ider in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect. Immun.* 70, 3371–3381. doi: 10.1128/IAI.70.7.3371-3381.2002

Salamatian, I., Ghaniei, A., Mosavari, N., Nourani, H., Keshavarz, R., and Eslampanah, M. (2020). Outbreak of avian mycobacteriosis in a commercial Turkey breeder flock. *Avian Pathol.* 49, 296–304. doi: 10.1080/03079457.2020.1740167

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

Seemann, T (2019). Snippy: Rapid Haploid Variant Calling and Core Genome Alignment. Available at: https://github.com/tseemann/snippy (Accessed May 19, 2020).

Shen, W., Le, S., Li, Y., and Hu, F. (2016). Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Slany, M., Ulmann, V., and Slana, I. (2016). Avian mycobacteriosis: still existing threat to humans. *Biomed. Res. Int.* 2016:4387461. doi: 10.1155/2016/4387461

Takahashi, H., Nakagawa, T., Yagi, T., Moriyama, M., Inagaki, T., Ichikawa, K., et al. (2015). Characterization of a novel plasmid, pmah135, from *mycobacterium avium* subsp. *Hominissuis*. *PLoS One* 10:e0117797. doi: 10.1371/journal.pone.0117797

Thirunavukkarasu, S., Plain, K. M., de Silva, K., Begg, D., Whittington, R. J., and Purdie, A. C. (2014). Expression of genes associated with cholesterol and lipid metabolism identified as a novel pathway in the early pathogenesis of *mycobacterium avium* subspecies *paratuberculosis*-infection in cattle. *Vet. Immunol. Immunopathol.* 160, 147–157. doi: 10.1016/j.vetimm.2014.04.002

Thorel, M.-F., Kichevsky, M., and Levy-Frebault, V. 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. Bacteriol.* 40, 254–260. doi: 10.1099/00207713-40-3-254

Timms, V. J., Daskalopoulos, G., Mitchell, H. M., and Neilan, B. A. (2016). The association of *mycobacterium avium* subsp. *Paratuberculosis* with inflammatory bowel disease. *PLoS One* 11:e0148731. doi: 10.1371/journal.pone.0148731

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

Tortoli, E., Meehan, C. J., Grottola, A., Fregni Serpini, G., Fabio, A., Trovato, A., et al. (2019). Genome-based taxonomic revision detects a number of synonymous taxa in the genus mycobacterium. *Infect. Genet. Evol.* 75:103983. doi: 10.1016/j.meegid.2019.103983

Tortoli, E., Rindi, L., Garcia, M., Chiaradonna, P., Dei, R., Garzelli, C., et al. (2004). Proposal to elevate the genetic variant mac-a, included in the mycobacterium avium complex, to species rank as *mycobacterium chimaera* sp. Nov. *Int. J. Syst. Evol. Microbiol.* 54, 1277–1285. doi: 10.1099/ijs.0.02777-0

Turenne, C. Y., Collins, D. M., Alexander, D. C., and Behr, M. A. (2008). *Mycobacterium avium* subsp. *Paratuberculosis* and *m. Avium subsp. Avium* are independently evolved pathogenic clones of a much broader group of *m. Avium* organisms. *J. Bacteriol.* 190, 2479–2487. doi: 10.1128/JB.01691-07

Turenne, C., Wallace, R., and Behr, M. (2007). Mycobacterium avium in the postgenomic era. *Clin. Microbiol. Rev.* 20, 205–229. doi: 10.1128/CMR.00036-06

Uchiya, K.-I., Takahashi, H., Yagi, T., Moriyama, M., Inagaki, T., Ichikawa, K., et al. (2013). Comparative genome analysis of *mycobacterium avium* revealed genetic diversity in strains that cause pulmonary and disseminated disease. *PLoS One* 8:e71831. doi: 10.1371/journal.pone.0071831 van Ingen, J., Boeree, M. J., Kösters, K., Wieland, A., Tortoli, E., Dekhuijzen, P. N. R., et al. (2009). Proposal to elevate mycobacterium avium complex its sequevar mac-q to *mycobacterium vulneris* sp. Nov. *Int. J. Syst. Evol. Microbiol.* 59, 2277–2282. doi: 10.1099/ijs.0.008854-0

van Ingen, J., Turenne, C. Y., Tortoli, E., Wallace, R. J. and Brown-Elliott, B. A. (2018). A definition of the mycobacterium avium complex for taxonomical and clinical purposes, a review'. *Int. J. Syst. Evol. Microbiol.* 68, 3666–3677.

Waddell, L. A., Rajic, A., Stark, K. D. C., and McEwen, S. A. (2015). The zoonotic potential of mycobacterium avium ssp paratuberculosis: a systematic review and meta-analyses of the evidence. *Epidemiol. Infect.* 143, 3135–3157. doi: 10.1017/S095026881500076X

Wang, J., Moolji, J., Dufort, A., Staffa, A., Domenech, P., Reed, M., et al. (2016). Iron acquisition in *mycobacterium avium* subsp. *Paratuberculosis*. *J. Bacteriol.* 198, 857–866. doi: 10.1128/JB.00922-15

Wang, J., Pritchard, J. R., Kreitmann, L., Montpetit, A., and Behr, M. A. (2014). Disruption of *mycobacterium avium* subsp. *Paratuberculosis*-specific genes impairs in vivo fitness. *BMC Genomics* 15:415. doi: 10.1186/1471-2164-15-415

Wang, W., Yang, J., Wu, X., Wan, B., Wang, H., Yu, F., et al. (2021). Difference in drug susceptibility distribution and clinical characteristics between *mycobacterium avium* and *mycobacterium intracellulare* lung diseases in Shanghai, China. *J. Med. Microbiol.* 70. doi: 10.1099/jmm.0.001358

Wayne, L. G., Good, R. C., Tsang, A., Butler, R., Dawson, D., Groothuis, D., et al. (1993). Serovar determination and molecular taxonomic correlation in *Mycobacterium avium*, Mycobacterium intracellulare, and *Mycobacterium scrofulaceum*: a cooperative study of the international working group on mycobacterial taxonomy. *Int. J. Syst. Bactreriol.* 43, 482–489. doi: 10.1099/00207713-43-3-482

Weigoldt, M., Meens, J., Bange, F.-C., Pich, A., Gerlach, G. F., and Goethe, R. (2013). Metabolic adaptation of *mycobacterium avium* subsp. *Paratuberculosis* to the gut environment. *Microbiology* 159, 380–391. doi: 10.1099/mic.0.062737-0

Whittington, R., Marsh, I., Choy, E., and Cousins, D. (1998). Polymorphisms in is1311, an insertion sequence common to *mycobacterium avium* and *m. Avium* subsp. *Paratuberculosis*, can be used to distinguish between and within these species. *Mol. Cell. Probes* 12, 349–358. doi: 10.1006/mcpr.1998.0194

Wibberg, D., Price-Carter, M., Rückert, C., Blom, J., and Möbius, P. (2020). Complete genome sequence of ovine *mycobacterium avium* subsp. *Paratuberculosis* strain jiii-386 (map-s/type iii) and its comparison to map-s/type i, map-c, and *M. avium* complex genomes. *Microorganisms* 9:70. doi: 10.3390/microorganisms9010070

Zhang, Y., Li, J., Li, B., Wang, J., and Liu, C. H. (2018). *Mycobacterium tuberculosis* mce3c promotes mycobacteria entry into macrophages through activation of β2 integrin-mediated signalling pathway. *Cell. Microbiol.* 20. doi: 10.1111/cmi.12800

**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.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 July 2022
doi: 10.3389/fmicb.2022.939406

# Genomic Insights Into the Interspecific Diversity and Evolution of *Mobiluncus*, a Pathogen Associated With Bacterial Vaginosis

*Yisong Li, Ying Wang and Jie Liu**

*School of Public Health, Qingdao University, Qingdao, China*

Bacterial vaginosis (BV) is a common vaginal infection and has been associated with increased risk for a wide array of health issues. BV is linked with a variety of heterogeneous pathogenic anaerobic bacteria, among which *Mobiluncus* is strongly associated with BV diagnosis. However, their genetic features, pathogenicity, interspecific diversity, and evolutionary characters have not been illustrated at genomic level. The current study performed phylogenomic and comparative genomic analyses of *Mobiluncus*. Phylogenomic analyses revealed remarkable phylogenetic distinctions among different species. Compared with *M. curtisii*, *M. mulieris* had a larger genome and pangenome size with more insertion sequences but less CRISPR-Cas systems. In addition, these two species were diverse in profile of virulence factors, but harbored similar antibiotic resistance genes. Statistically different functional genome profiles between strains from the two species were determined, as well as correlations of some functional genes/pathways with putative pathogenicity. We also showed that high levels of horizontal gene transfer might be an important strategy for species diversification and pathogenicity. Collectively, this study provides the first genome sequence level description of *Mobiluncus*, and may shed light on its virulence/pathogenicity, functional diversification, and evolutionary dynamics. Our study could facilitate the further investigations of this important pathogen, and might improve the future treatment of BV.

**Keywords: *Mobiluncus*, comparative genomics, interspecific divergence, pathogenicity, horizontal gene transfer, bacterial vaginosis**

OPEN ACCESS

***Edited by:***
*Yufeng Wang,*
*University of Texas at San Antonio, United States*

***Reviewed by:***
*Wei Ke,*
*Rutgers, The State University of New Jersey, United States*
*Liang Wang,*
*Guangdong Provincial People's Hospital, China*

****Correspondence:***
*Jie Liu*
*jl5yj@hotmail.com*

***Specialty section:***
*This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology*

***Received:*** *09 May 2022*
***Accepted:*** *16 June 2022*
***Published:*** *05 July 2022*

***Citation:***
*Li Y, Wang Y and Liu J (2022) Genomic Insights Into the Interspecific Diversity and Evolution of Mobiluncus, a Pathogen Associated With Bacterial Vaginosis. Front. Microbiol. 13:939406. doi: 10.3389/fmicb.2022.939406*

## INTRODUCTION

Bacterial vaginosis (BV), a common gynecological disease characterized by vaginal discharge, affects roughly a quarter of women worldwide and costs an estimated $4.8 billion annually (Sobel, 2000; Javed et al., 2019; Peebles et al., 2019). BV has been associated with an increased risk of various health problems, including sexually transmitted infections, adverse pregnancy outcomes (e.g., preterm births, premature rupture of membranes), pelvic inflammatory disease, increased susceptibility to HIV infection, and other chronic health issues (Taha et al., 1998; Schwebke and Desmond, 2005; Onderdonk et al., 2016). The pathogenesis of BV is still a subject of debate (Cherpes et al., 2008; Muzny and Schwebke, 2016), whereas it has been determined that BV is characterized by a decrease in the levels of *Lactobacilli* and an overgrowth of opportunistic bacteria including anaerobes or microaerophiles such as *Prevotella*, *Gardnerella*, and *Mobiluncus* genera (Hillier, 1993; Sha et al., 2005). Among these, the abundance of Gram-negatively stained and curved rod-shaped bacteria, represented by *Mobiluncus*, has been considered as one of the key indicators of Nugent score, the "gold standard" for BV diagnosis (Nugent et al., 1991; Srinivasan et al., 2013).

*Mobiluncus* organisms were initially recognized in vaginal fluid as early as 1895 and were first isolated in 1913 (Curtis, 1913). Nowadays, much interest has revolved around *Mobiluncus* since women with higher Nugent scores (predominantly because of morphotypes consistent with *Mobiluncus*) are more likely to fail therapy than those with lower scores (Schwebke and Desmond, 2007), and the presence and persistence of *Mobiluncus* spp. was found to be highly associated with recurrence of BV (Meltzer et al., 2008). In addition, the production of malic acid and trimethylamine by *Mobiluncus* strains have been reported to give rise to vaginal irritation and unpleasant odor (Africa et al., 2014). For these reasons, more efforts have been made to characterize their resistance mechanisms and virulence factors (Spiegel, 1987; Zeng et al., 2020; Zhang et al., 2020). Nevertheless, the definite role of *Mobiluncus* in BV pathogenesis still remains largely elusive.

Analysis of the 16S rRNA gene sequences has revealed that *Mobiluncus* genus mainly contains two distantly related species, *M. curtisii* and *M. mulieris*. Previous reports from a number of laboratories have shown that these two species can be differentiated based on physical and biochemical properties. It has been demonstrated that *M. curtisii* and *M. mulieris* comprise short curved and long straight (or slightly curved) rods (Hoyles et al., 2004; Onderdonk et al., 2016), respectively, and show variation in growth in different liquid media (Taylor-Robinson and Taylor-Robinson, 2002). In addition, antigenic profiles of the two species are also distinct (Roberts et al., 1985; Gatti et al., 1997), and *M. mulieris* can stimulate a TLR5-mediated response in host, while *M. curtisii* cannot (Dela Cruz et al., 2021). Antimicrobial susceptibility and clonality of *Mobiluncus* also vary widely among species and even strains (Spiegel, 1987; Zhang et al., 2020). Recently, a new species collected from pig gut, namely *M. porci*, has been described (Wylensek et al., 2020). However, these studies were mostly based on phenotypic data, while genomic features, including genetic diversity and evolutionary history, of/between *Mobiluncus* species have not been clearly elucidated yet.

In the present study, we carried out an in-depth comparative genomic analysis of 38 publicly available genomic sequences of *Mobiluncus*, aiming to investigate the genomic diversity of this taxon. We compared the status of various virulence and antibiotic resistance genes (ARGs) among the strains in order to unleash the potential underlying mechanism of pathogenicity and resistance. Moreover, we identified genes that may contribute to the differentiation between species, and tried to build linkages between genetic differences (gene functions and metabolic pathways) and potential pathogenicity. Finally, we uncovered the evolutionary events that may contribute to these variabilities, especially the horizontal gene transfer (HGT) events. Altogether, our study not only provides first insights into genomic features and evolution of the genus *Mobiluncus*, but also has implications for improved understanding of the pathogenic mechanism and putative treatment of this pathogen.

## MATERIALS AND METHODS

### Genome Data Set

Contigs or scaffolds of the genome sequences for the members of the genus *Mobiluncus* were downloaded from the NCBI genomes FTP site (April 2021)[1]. To avoid bias, genomes with estimated contamination > 5% or completion < 95% were excluded based on CheckM results (Parks et al., 2015). Taxonomy assignment of these genomes was performed by the Genome Taxonomy Database (GTDB) toolkit (Chaumeil et al., 2019) and based on LPSN (List of Prokaryotic names with Standing in Nomenclature) database (Parte, 2018). Contigs of different species were reordered using the Move Contig tool in Mauve software (Darling et al., 2010) against the complete genome of the *M. curtisii* ATCC 43063 and *M. mulieris* DSM 2710, respectively. Pairwise genome alignment was carried out by the lastz program[2], and the results were visualized using AliTV (Ankenbrand et al., 2017). Pairwise whole genome average nucleotide identity (ANI) values were computed by FastANI (Jain et al., 2018).

### Genome Annotation

All genomes were reannotated using Prokka with default settings (Seemann, 2014). Functional annotation and classification of proteins were performed by sequence comparison using DIAMOND BLASTP (*E*-value 1e-05, coverage 0.5, and identity 40%) (Buchfink et al., 2015) against the recently updated Clusters of Orthologous Group (COG) database (Galperin et al., 2021). KEGG Automated Annotation Server (KAAS) (Moriya et al., 2007) was used for pathway mapping of species-specific genes. Insertion sequences (ISs) were identified by BLASTN against the ISFinder database (*E*-value 1e-05) (Siguier et al., 2006). The prediction of clustered regularly interspaced short palindromic repeat (CRISPR) in the genome was assessed by the CRISPRCasFinder tool (Couvin et al., 2018), and only CRISPRs classified with evidence levels 3 and 4 were considered. Potential ARGs and putative virulence factors (VFs) encoded in genomes were identified through BLASTP searches of the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2020) and the Virulence Factors Database (VFDB) (Liu et al., 2019), respectively. The most differentiating COG entries between *M. curtisii* and *M. mulieris* were determined by SIMPER analysis by using "simper" function of the vegan R package[3]. "ordinate" function in the phyloseq R package (McMurdie and Holmes, 2013) was used to determine the variation in the functional profiles of HGT genes with non-metric multidimensional scaling (NMDS) analysis on BrayCurtis dissimilarity matrices. Pairwise

[1] https://www.ncbi.nlm.nih.gov/genomes

[2] https://github.com/lastz/lastz

[3] https://github.com/vegandevs/vegan

comparisons of species-specific genomic regions were visualized by EasyFig software (Sullivan et al., 2011).

## Pan-Genome and Phylogenomic Analyses

Homologous gene families were calculated using GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013) with the OrthoMCL clustering algorithm, and cloud, shell, and (soft-) core pangenome components were also derived. Pan-genome statistics were computed by PanGP (Zhao et al., 2014). Marker-based phylogenetic tree of the genus *Mobiluncus* was constructed by the GET_PHYLOMARKERS pipelines (Vinuesa et al., 2018) running in default mode based on nucleotide sequences of 329 high-quality marker genes. In addition. an absence/presence (0/1) matrix of dispensable genes was built according to GET_HOMOLOGUES results, and was subjected to R package pvclust (Suzuki and Shimodaira, 2006) for hierarchical clustering analysis with 1,000 bootstrap replicates with two types of *p*-values: AU (approximately unbiased) *p*-value and BP (bootstrap probability) value.

## Identification of Potential Horizontal Genes

In order to predict the gain and loss of each homologous gene family across ancestral nodes during the evolution of *Mobiluncus*, the pangenome matrix and the rooted phylogenomic tree were used as inputs for COUNT software (Csűrös, 2010) to calculate posterior probabilities (cut-off was set at 70%). We also used HGTector software (Zhu et al., 2014) to detect genes in each genome that were potentially acquired through HGT. During this process, quality cutoffs for DIAMOND BLASTP results were *E*-value ≤ 1e-05, sequence identity ≥ 50%, and coverage of query sequence ≥ 50%. *Mobiluncus* (rank, genus; taxon identifier 2050) was set as the *self* group, and *Actinomycetaceae* (rank, family; taxon identifier 2049) was set as the *close* group.

# RESULTS

## Phylogeny and Genome Overview of the Genus *Mobiluncus*

All 40 *Mobiluncus* genomes were downloaded from GenBank database (April 2021). Two genomes (strain FDAARGOS_303 and strain NCTC11820) were filtered out as they contained more contamination (8.53 and 11.87%, respectively). Therefore, a total of 38 high-quality *Mobiluncus* genomes were analyzed in this study, including four complete and 34 draft genome sequences (**Supplementary Table 1**). Based on the similarity observed by GTDB-tk, two valid species are present within the genus, with 18 genomes belonging to *M. curtisii* and 19 genomes belonging to *M. mulieris*. All of these strains were isolated from human vagina (except for two with missing data). In addition, based on LPSN database, strain RF-GAM-744-WT-7 (isolated from pig feces) was classified as *M. porci* (Wylensek et al., 2020). To further elucidate their genetic relatedness, a genome-wide ANI plot was generated (**Figure 1A**). The intraspecies ANI values of *M. curtisii* and *M. mulieris* were higher than 95.7 and 97.4%, respectively, which exceeded the recommended 95% threshold value for intraspecific prokaryotic strains (Richter and Rossello-Mora, 2009), while the inter-species values were lower than 78.4%. To further evaluate the intra-genus differentiation and evolutionary relationships within the genus, phylogenomic reconstruction was performed based on 329 high-quality phylogenomic marker (core) genes (**Figure 2A**). The resulting tree shows two major clades corresponding to *M. curtisii* and *M. mulieris* species, and one distinct branch formed by the single strain *M. porci* RF-GAM-744-WT-7 clustering with *M. mulieris*. A hierarchical clustering tree based on the content of dispensable genes showed similar topological structure (**Figure 2B**), but with *M. porci* RF-GAM-744-WT-7 located on the outskirt of *M. curtisii*. Given the representativeness, the following analyses will focus more on the two validly published BV-associated species, *M. curtisii* and *M. mulieris*.

The characteristics of the genomes studied here are shown in **Figures 1B–H** and **Supplementary Table 1**. The average genome size of *M. curtisii* was 2.17 Mbp and was significantly less than *M. mulieris* at 2.53 Mbp (Wilcoxon test, $p < 0.001$). Consequently, *M. curtisii* on average contained fewer protein coding genes at 1,844 compared to *M. mulieris* at 2,213 (Wilcoxon test, $p < 0.001$), and between species showed collinearity with abundant gene arrangement. GC content levels were similar between species, with an average of 55.3%, although *M. curtisii* had a slightly larger GC content than *M. mulieris* (mean, 55.5 and 55.0%, respectively; Wilcoxon test, $p < 0.001$). Strikingly, 11 of 18 *M. curtisii* strains contained two copies of 16S rRNA genes, while only 4 of 19 for *M. mulieris*, and the remaining strains of the two species contained only one copy (Fisher exact test, $p = 0.0201$). The average number of ISs per genome was 10.8 in *M. curtisii* while 17.8 in *M. mulieris*, with *M. mulieris* strain NCTC11497 harboring the greatest number ($n = 68$). CRISPR-Cas system presented in nearly all (17 of 18) *M. curtisii* strains (16 TypeIE and 1 TypeIIC), but only 11 of 19 for *M. mulieris* (4 TypeIE and 7 TypeIIC) (Fisher exact test, $p = 0.0188$). Taken together, these genomic features suggested an apparent genomic divergence between *M. curtisii* and *M. mulieris*.

## Antibiotic Resistance Genes and Potential Virulence Factors of *Mobiluncus*

A total of 26 distinctive putative ARGs were identified. Each genome contained 14.02 ARGs averagely, and no remarkable difference in the total number or profiles of ARGs was found between *M. mulieris* and *M. curtisii* (**Figures 1I**, **3A**). Four ARGs were shared by all *Mobiluncus* strains, including *rpoB* and *rpoB2* (ARO:3004480 and ARO:3000501, respectively; both conferring resistance to rifampicin), *bcrA* (ARO:3002987, conferring bacitracin resistance) and *mtrA* (ARO:3000816, encoding a transcriptional activator of the MtrCDE multidrug efflux pump). Specifically, *otrC* (ARO:3002894) was found to be *M. curtisii*-specific, which encoded a tetracycline resistance efflux pump.

We next investigated the distribution of putative virulence genes. Overall, genomes of *M. mulieris* contained more VFs

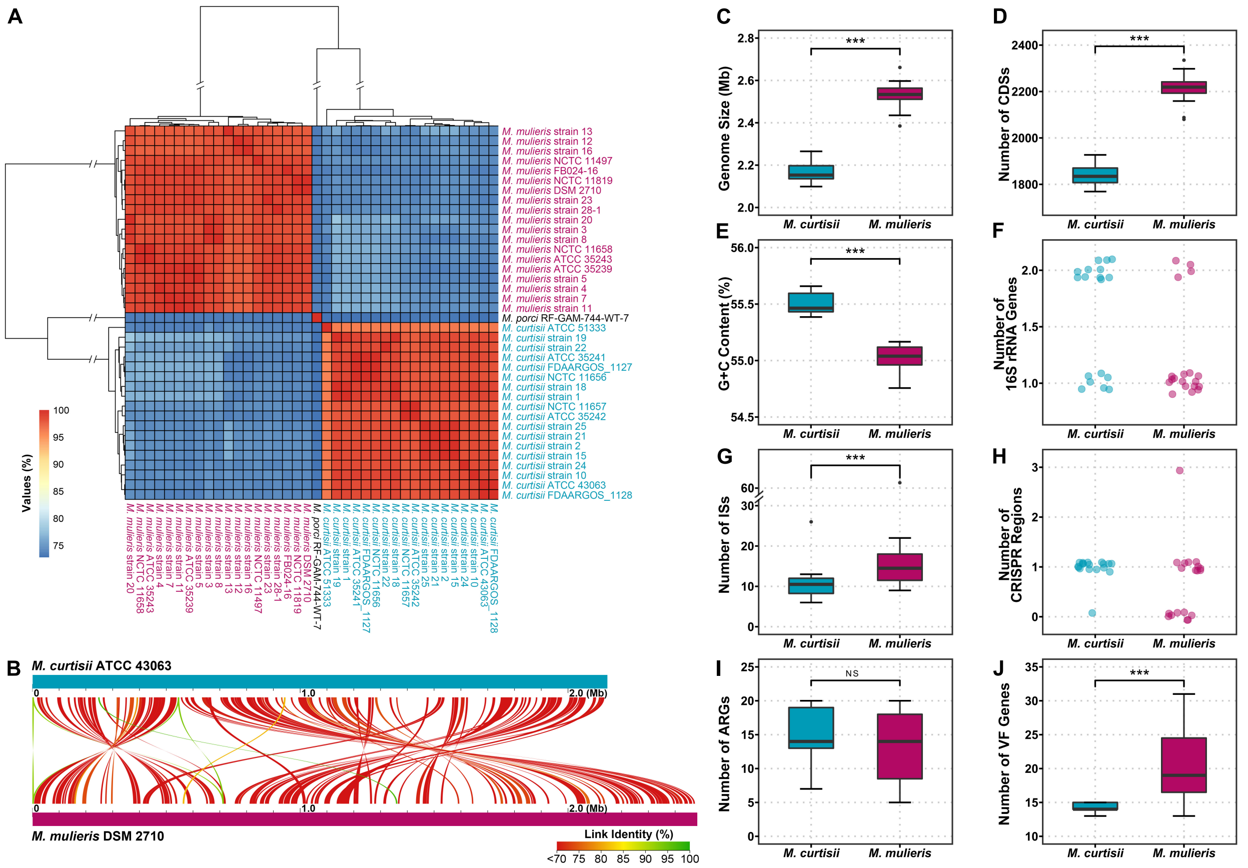


**FIGURE 1 |** Genomic features of the genus *Mobiluncus*. **(A)** Hierarchical clustering based on ANI values of 38 *Mobiluncus* strains. **(B)** Whole genome alignment of *M. curtisii* and *M. mulieris*. Color denotes percent similarity of links (see legend). **(C–J)** Comparison of genomic characteristics between *M. curtisii* and *M. mulieris*. The boxplot shows the median, and the first and third quartiles as the lower and upper hinges. Outliers are indicated as dots. Asterisks (***) indicate significant differences (significance level of 0.001, Wilcoxon test). NS, not significant; ISs, insertion sequences; CRISPR, clustered regularly interspaced short palindromic repeat; ARGs, antibiotic resistance genes; VF, virulence factor.

(mean, 20.42; median, 19) than those of *M. curtisii* (mean, 14.28; median, 14) (**Figure 1J**), and the profiles of the VFs could largely differentiate the two species (**Figure 3B**). Ten VFs were shared by all strains, including stress adaptation (*clpC*, *clpE,* and *clpP*), regulation (*phoP* and *relA*), adherence (*groEL* and *tufA*), secretion (*lirB*), immune evasion (*wbtL*) and others (*htpB*), suggesting they might play key roles in pathogenicity of *Mobiluncus* strains. In addition, six VFs were prevalent in *M. mulieris* but absent in *M. curtisii* strains. Among these, five genes (*flaABCDE*) are associated with bacterial flagellin proteins, which serve as mediators of pathogenicity and host immune responses (Ramos et al., 2004); and the last gene is *ideR*, protein of which has been reported to be the key regulator of VFs and iron homeostasis in *Mycobacterium tuberculosis* (Pandey and Rodriguez, 2014). Moreover, there were two VFs only present in all *M. curtisii* strains: one is *galE*, encoding a UDP-galactose-4-epimerase involved in the biosynthesis of capsular or O-antigen polysaccharide units in many bacterial pathogens (Agarwal et al., 2007; Li et al., 2014); and the other is *pscN*, which encoded ATPase of Type III secretion system as a main VF reported in *Pseudomonas aeruginosa* (Lee and Rietsch, 2015; Ngo et al., 2020).

## Interspecific Pangenome Variation of *Mobiluncus*

To explore the interspecific pangenome variation, we characterized the core and pan-genomes of *M. curtisii* and *M. mulieris* separately (**Figure 4**). The pangenome of *M. curtisii* contained 2,576 genes, whereas that of *M. mulieris* contained 3,507 genes. There was little difference in the core genome size (1,540 and 1,539 genes, respectively; softcore: 1,598 and 1,619 genes, respectively) between the two species. However, it revealed that the pangenome of *M. curtisii* comprised 450 cloud genes (accounting for 17.47% of the total genes) and 528 shell genes (20.50%), much less than those of *M. mulieris* (814 and 23.21%, 1,074 and 30.62%, respectively). According to a power-law regression, both species pangenomes were "open", with $B_{pan}$ = 0.19 (*M. curtisii*) and 0.15 (*M. mulieris*). Taken together, both pangenomes appeared to be boundless, while that of the *M. mulieris* was relatively more extensive and heterogeneous.

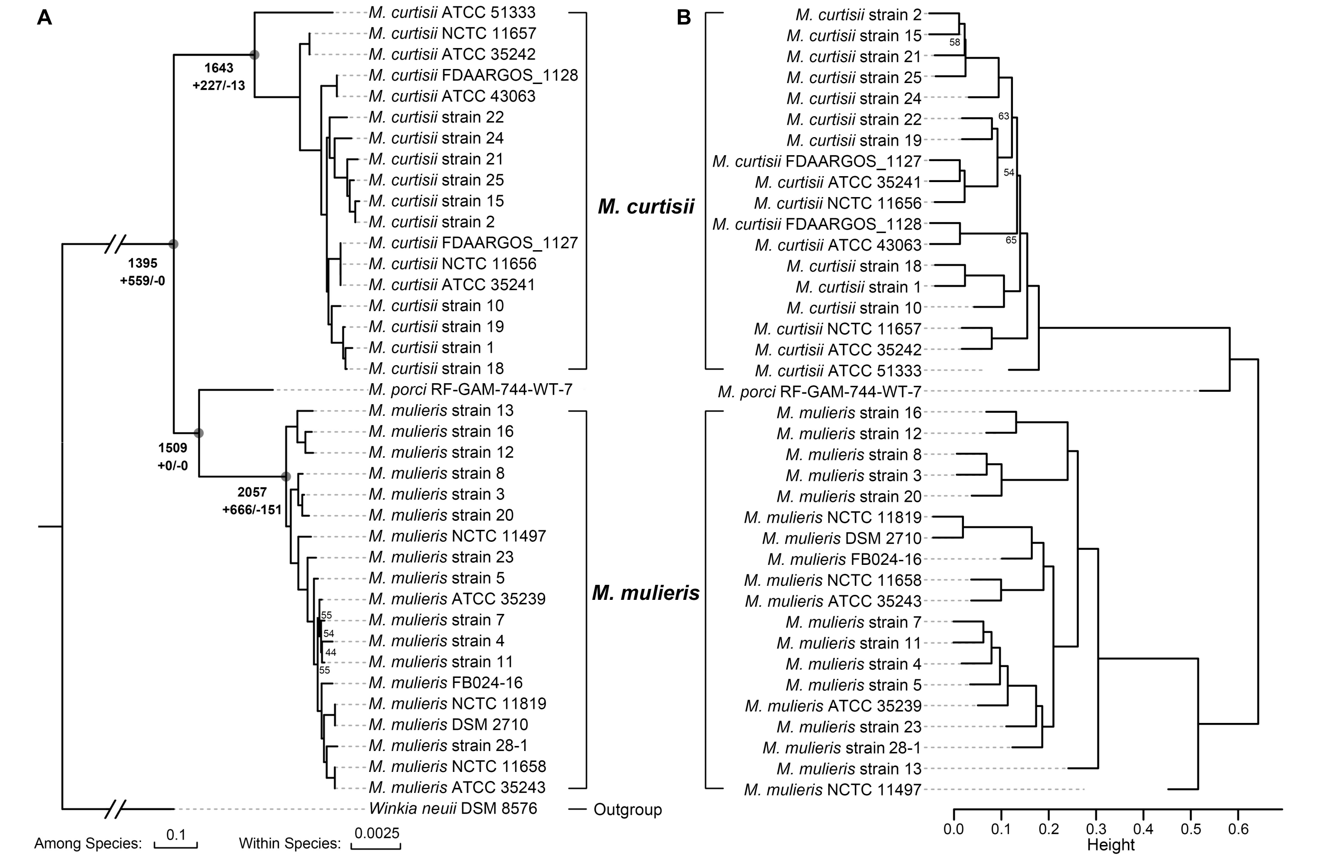


**FIGURE 2 |** Phylogeny of the genus *Mobiluncus*. **(A)** Maximum-likelihood phylogenomic tree. The tree was constructed based on 329 core genes that could give a well-resolved tree topology without recombination. The total numbers of ancestral orthologous genes present at each of the nodes are shown next to the deep nodes, and the numbers of gene gain (+) and loss (–) events are indicated below. Bootstrap values less than 70% are shown at the nodes. The scale bars indicate 10 and 0.25% sequence divergences among and within species, respectively. **(B)** Hierarchical cluster analysis based on the presence or absence of dispensable genes. Height indicates the dissimilarity between genomes. BP (bootstrap probability) values less than 70% from 1,000 replicates are shown, and all AU (approximately unbiased) *p*-values are > 70%.

## Functional Divergence Between *Mobiluncus* Species

To investigate functional differentiation between *M. curtisii* and *M. mulieris*, we first explored the COG functional classification for all genes in each genome (**Figure 5A**). As a result, *M. curtisii* strains had a higher proportion of genes classified in COG categories E (amino acid transport and metabolism), H (coenzyme transport and metabolism), M (cell wall/membrane/envelope biogenesis) and P (inorganic ion transport and metabolism), while *M. mulieris* strains was significantly enriched for genes classified in COG categories G (carbohydrate transport and metabolism) and V (defense mechanisms). In addition, a total of 18 COGs were detected that significantly contributed most to the dissimilarity between species (SIMPER analysis, > 0.1% contribution, $p < 0.01$) (**Figure 5B** and **Supplementary Table 2**), with two-thirds were related to COGs G, V and M. For example, genes associated with cell wall binding/biosynthesis (COG2247 and COG0463) and lipoprotein transport (COG4591) are more abundant in *M. curtisii*, while *M. mulieris* strains contains more genes from RelBE/YafQ-DinJ/Txe-Axe toxin-antitoxin module (COG3077, COG2026, COG4115, and COG3041) and transmembrane proteins of sn-glycerol-3-phosphate transport system (COG0395 and COG1175).

To further investigate functional differentiation between the species, we explored the species-specific genes that are universal (> 90%) in one species but absent in the other. We found 385 and 429 orthologous genes (OGs) that were specific to *M. curtisii* and *M. mulieris*, respectively (**Supplementary Table 3**). Some of the OGs were located physically adjacent and clustered into genomic regions (**Figure 6**), which might perform certain complicated or special roles in extending the metabolic/pathogenic pathways. One such region was composed of several arginine biosynthetic genes. Interestingly, besides operon *argDRGH*, within the region genomes of *M. curtisii* also contained genes of *argCJB*, whereas *M. mulieris* lacked but instead harbored operon *carAB* elsewhere, which has been reported to be necessary for pyrimidine nucleotide and arginine biosynthesis (Han and Turnbough, 1998). Another region

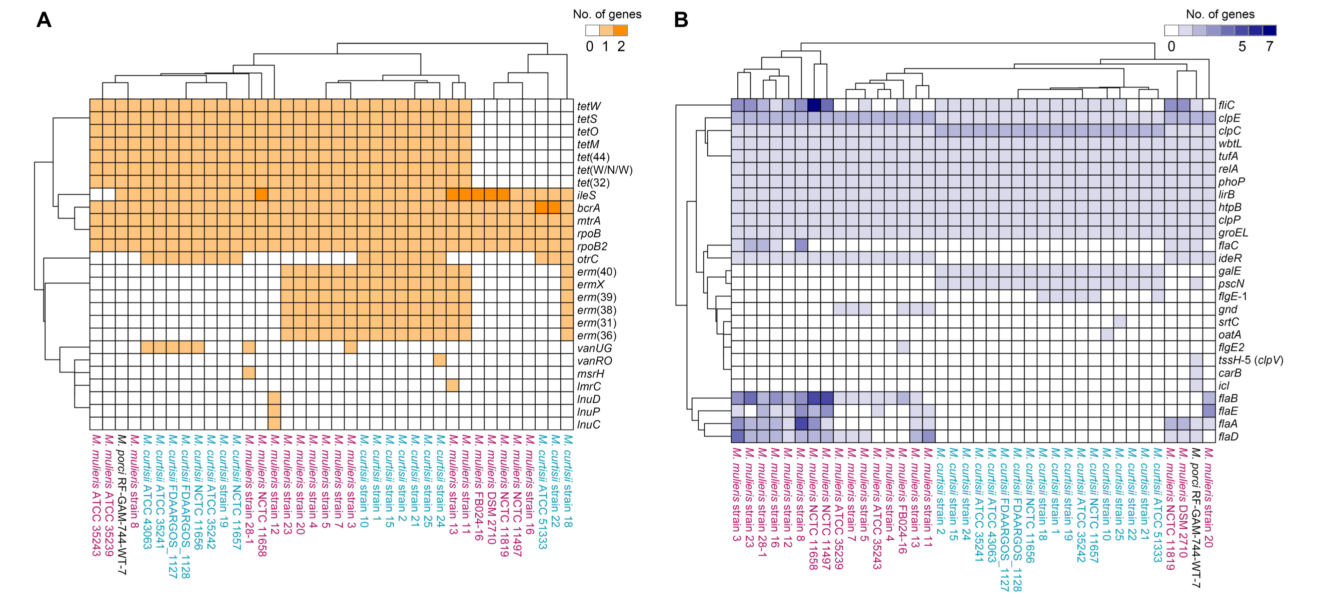


**FIGURE 3 |** Hierarchically clustered heatmaps of the distributions of the putative antibiotic resistance genes **(A)** and virulence factors **(B)** in *Mobiluncus* genomes.

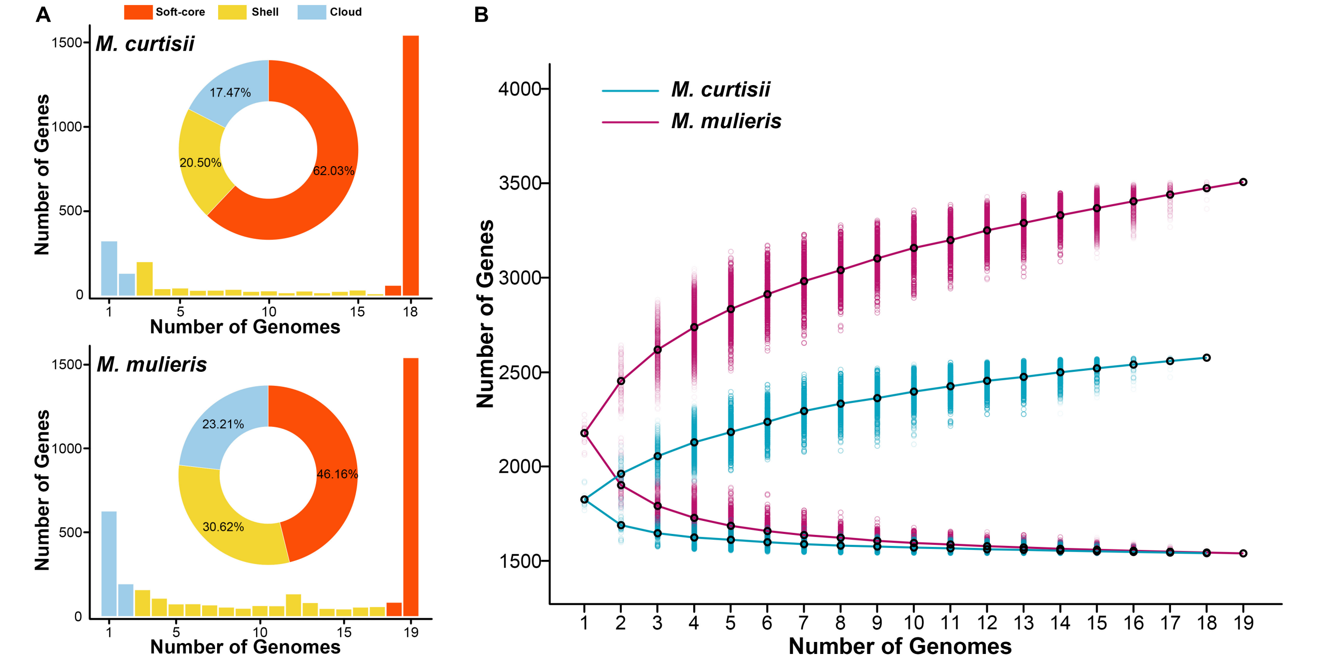
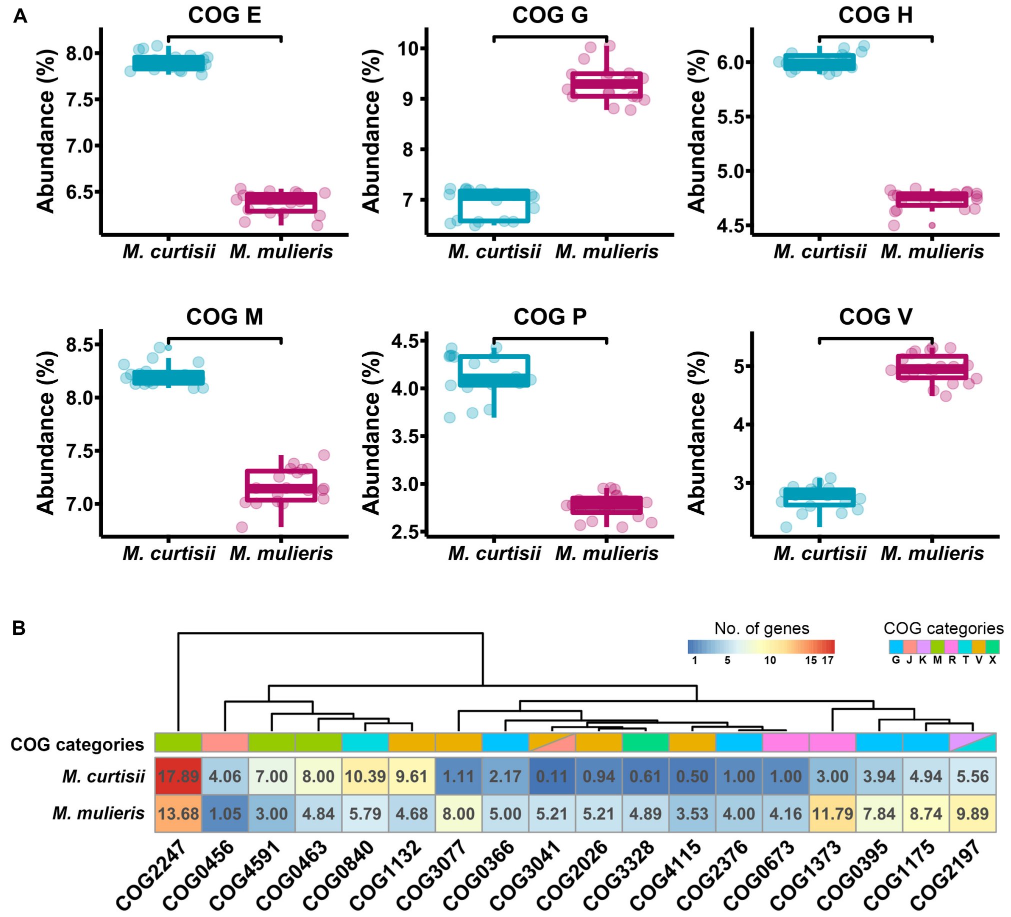


**FIGURE 4 |** Pangenome summary statistics of *M. curtisii* and *M. mulieris*. **(A)** Histogram distributions of soft-core, shell, and cloud genes. Pie chart displays percentages of each part of the total genes. **(B)** The sizes of pan- and core-genomes in relation to numbers of genomes added into the gene pool.

contained two pathways, one was involved in the molybdopterin biosynthesis, encoding an ABC-type molybdate transport system and a biosynthetic gene cluster of molybdenum cofactor (MoCo), and another associated with nitrate respiration, encoding a nitrate reductase operon *narKGHJI* (only present in 73.7% *M. curtisii* strains). Moreover, although all genomes had a series of genes related to histidine biosynthesis, another eight *his* genes (*hisF*, *hisI*, *hisG*, *hisA*, *hisH*, *hisB*, *hisC,* and *hisD*) were unique to *M. curtisii* strains. Similarly, operon *nadABC* existed only in *M. curtisii*, enabling them biosynthesize NAD$^+$ in both the salvage and the *de novo* pathways. Another *M. curtisii*-specific region contained genes of LIV system, which is responsible for the transport of branched-chain amino acids, such as leucine, isoleucine, and valine (Adams et al., 1990).



| | COG2247 | COG0456 | COG4591 | COG0463 | COG0840 | COG1132 | COG3077 | COG0366 | COG3041 | COG2026 | COG3328 | COG4115 | COG2376 | COG0673 | COG1373 | COG0395 | COG1175 | COG2197 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *M. curtisii* | 17.89 | 4.06 | 7.00 | 8.00 | 10.39 | 9.61 | 1.11 | 2.17 | 0.11 | 0.94 | 0.61 | 0.50 | 1.00 | 1.00 | 3.00 | 3.94 | 4.94 | 5.56 |
| *M. mulieris* | 13.68 | 1.05 | 3.00 | 4.84 | 5.79 | 4.68 | 8.00 | 5.00 | 5.21 | 5.21 | 4.89 | 3.53 | 4.00 | 4.16 | 11.79 | 7.84 | 8.74 | 9.89 |

**FIGURE 5 |** Functional divergences between *M. curtisii* and *M. mulieris*. **(A)** Boxplot of abundance of differential COG categories. All the distributions were significantly different (Wilcoxon test, $p$ < 10-5). **(B)** Heatmap of COGs that significantly contributed most to the dissimilarity between *M. curtisii* and *M. mulieris*. Different COG categories are shown in different colors.

We also detected three *M. mulieris*-specific genomic regions. One region was composed of genes of an ABC-type guanosine uptake system NupNOPQ, which inserted in an operon *deoD-deoA-cdd-manB-DeoC*, an unusual *deoCABD*-like operon perhaps serving as deoxynucleotide degradation also found in *Mollicutes* and other pathogens (Christensen et al., 2003; Bizarro and Schuck, 2007). The second was composed of a complete sialic acid catabolic gene cluster *nanAKE*. Interestingly, although *M. curtisii* did not contain the cluster, a sialidase gene (*nanH*) was found to be *M. curtisii*-specific. We also detected a *lldPEFG* operon (orthologs of *lutABC* and *lctP*) in *M. mulieris*, which was implicated in lactate utilization and has been reported to be involved in biofilm formation and pathogenesis in many pathogenic bacteria (Chai et al., 2009; Jiang et al., 2014).

## Gain and Loss of Genes During the Evolution of *Mobiluncus*

To decipher the evolutionary history of the genus *Mobiluncus*, we first assessed the gain and loss events that have occurred of ancestral nodes of species on the phylogenomic tree (**Figure 2A**). The last common ancestor of the genus *Mobiluncus* was inferred to possess 1,395 gene families. Both of the *M. curtisii* and *M. mulieris* genomes have experienced a massive expansion, with 227 and 666 gene gains have been identified occurred at the divergence of the two species, with only 13 and 151 gene losses, respectively. Next, we further examined the potential horizontal genes in *Mobiluncus* genomes and tracked the potential donor. As a result, a total of 5,000 predicted HGT events were identified, with an average genome containing 131.58 (median, 130) horizontally transferred genes. Interestingly, although *M. mulieris* has a bigger genome than *M. curtisii*, predicted HGT events showed no significant variation between the two species (mean: 131.53 and 129.78, respectively). This may be because the transferred genes of *M. mulieris* comprised more shell and cloud genes (21.4%), while only 9.9% for *M. curtisii*. Correspondingly, 90.1% of the transferred genes of *M. curtisii* were soft-core genes, while only 78.6% for *M. mulieris* (**Figure 7A**). This content variation also presented at the gene family level.

We next tracked the potential donor of the potential horizontal genes in *Mobiluncus* genomes. Among the 5,000 identified potential HGTs, 57.7% of the donors could be

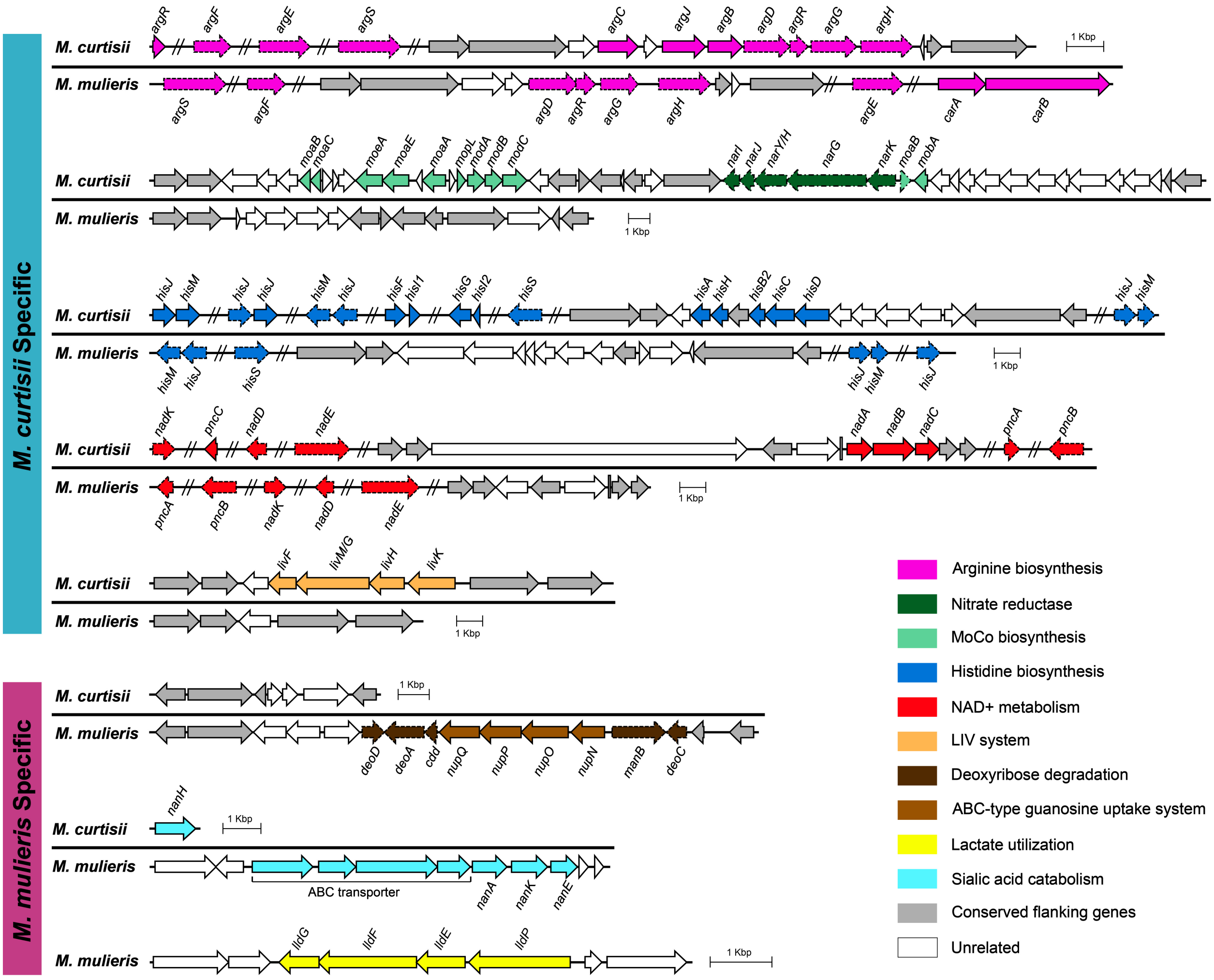

M. curtisii Specific
M. mulieris Specific
M. curtisii
M. mulieris
argR argF argE argS argC argJ argB argD argR argG argH
argS argF argD argR argG argH argE carA carB
moaB moaC moeA moeE moaA mopL modA modB modC narI narJ narY/H narG narK moaB mobA
hisJ hisM hisJ hisJ hisM hisJ hisF hisI1 hisG hisI2 hisS hisA hisH hisB2 hisC hisD hisJ hisM
hisM hisJ hisS hisJ hisM hisJ
nadK pncC nadD nadE nadA nadB nadC pncA pncB
pncA pncB nadK nadD nadE
livF livM/G livH livK
deoD deoA cdd nupQ nupP nupO nupN manB deoC
nanH
ABC transporter
nanA nanK nanE
lldG lldF lldE lldP
1 Kbp
Arginine biosynthesis
Nitrate reductase
MoCo biosynthesis
Histidine biosynthesis
NAD+ metabolism
LIV system
Deoxyribose degradation
ABC-type guanosine uptake system
Lactate utilization
Sialic acid catabolism
Conserved flanking genes
Unrelated

**FIGURE 6 |** Genetic organizations of the regions containing species-specific genes. Only genes that are universal (> 90%) in one species but absent in the other are determined as species-specific genes. Conserved genomic regions flanking these genes are colored in gray. Non-specific genes are marked with dashed lines. MoCo, molybdenum cofactor.

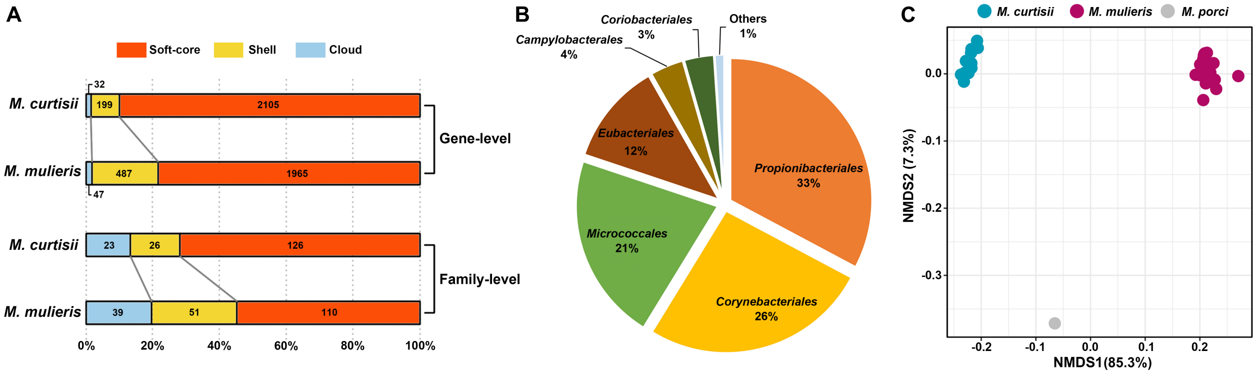


**FIGURE 7 |** Analysis of predicted transferred genes. **(A)** Distribution of the predicted transferred genes in soft-core, shell, and cloud genes. Numbers of transferred genes and gene families are indicated. **(B)** Pie chart displays percentages of different donors of HGT events at order level. **(C)** Non-metric multidimensional scaling (NMDS) plot of COG entries showing distinct clustering of each species.

annotated at the phylum level, of which 94.7% were from *Actinobacteria*. The other phyla included *Firmicutes* (4.51%), *Proteobacteria* (0.76%) and *Chloroflexi* (0.035%). In addition, potential donor taxa of 473 genes could be identified at the order level (**Figure 7B**). The orders *Propionibacteriales*, *Corynebacteriales,* and *Micrococcales* appeared to be the main donor taxa, accounting for 80% of the total cross-order HGT genes, while order *Eubacteriales* was the most non-actinobacteria donor. We also revealed a different functional profile of putative transferred genes between species. *M. curtisii* acquired a higher proportion of genes classified in COG E (amino acid transport and metabolism), while *M. mulieris* was biased toward COG categories J (translation, ribosomal structure and biogenesis) and V (defense mechanisms) (**Supplementary Figure 1**). This result was partially consistent with the functional divergence between *Mobiluncus* species described above (**Figure 5A**). Meanwhile, based on the number of proteins annotated to each COG entries, the three species showed different functional profiles of the HGT genes (PERMANOVA test, $p = 0.001$; **Figure 7C**), implying HGT contributed to the functional divergence of *Mobiluncus* species.

## DISCUSSION

More than 60 years have passed since BV first described (Gardner and Dukes, 1955), and even now, its etiology and the reason for global prevalence remain unclear (Kenyon et al., 2013; Coudray and Madhivanan, 2020). Common opportunistic bacteria causing BV include *Prevotella*, *Gardnerella vaginalis* and *Mobiluncus* (Thorsen et al., 1998; Coudray and Madhivanan, 2020), and the abundance of *Mobiluncus* strains always represents a higher Nugent score and a higher possibility to fail therapy (Schwebke and Desmond, 2007; Meltzer et al., 2008). For the first time, in the current study we tried to reveal genomic details of the genus *Mobiluncus*, to gain more insights into the genomic features, VF and ARG profiles, functional repertoire and the evolutionary history of *Mobiluncus* diversification. Such information would provide theoretical foundation for further studies on the pathogenicity, therapy and discrimination of *Mobiluncus* species.

Efforts have been made to distinguish the two main species of *Mobiluncus* on the basis of morphological and biochemical differences, especially on the antigenic profiles (Roberts et al., 1984, 1985; Spiegel, 1987; Zhang et al., 2020). In this study, we performed a robust phylogenomic reconstruction to verify the degree of differentiation among species, emphasizing the genetic differences between *M. curtisii* and *M. mulieris*. We showed that the genome size of *M. mulieris* was relatively larger, with more gene family gains predicted across its evolution and a more open pangenome. This is consistent with the fact that *M. mulieris* strains comprised more ISs but less CRISPR-Cas systems within the genomes, both of which have been reported to play important roles in the bacterial genome instability (Darmon and Leach, 2014; Hatoum-Aslan and Marraffini, 2014). In addition, a genomic investigation on ARG and VF profiles showed four ARGs and ten VFs were found to be prevalent in all *Mobiluncus* strains, while the remaining other genes exhibited sporadic distribution patterns. Moreover, VFs profiles were able to distinguish *M. curtisii* from *M. mulieris*, whereas ARG profiles were not. Correspondingly, previous experimental studies have also revealed significant intra- and inter-species heterogeneity of antimicrobial susceptibility (Spiegel, 1987; Zhang et al., 2020). Also, it should be mentioned that although the role of *Mobiluncus* in the etiology and pathology of BV remains unclear, these two species may exhibit different pathogenicity and distribution during the disease process (Meltzer et al., 2008; Onderdonk et al., 2016; Arries and Ferrieri, 2022), sometimes even contradictory (Schwebke and Lawing, 2001; Salinas et al., 2020). Nevertheless, the VF and ARG profiles revealed in this work may provide guidance for the future treatment of *Mobiluncus* infection.

We also detected a series of metabolic pathways that showed apparent species specificity, most of which have been reported to be associated with virulence and adoption of pathogenic organisms. For example, the role of arginine biosynthesis in virulence has been reported to be crucial for full virulence of *Aspergillus fumigatus* in insects (Dietl et al., 2020), and we have showed that the pathway of arginine biosynthesis in *M. curtisii* and *M. mulieris* was different, perhaps suggesting a different utilizing efficiency. Furthermore, the capabilities of MoCo biosynthesis and nitrate reduction were only found in *M. curtisii* strains, both of which have been implicated in pathogenesis of a number of bacterial infections (Williams et al., 2011; Andreae et al., 2014; Almeida et al., 2017); more *his* genes were found in *M. curtisii* strains, perhaps enabling them capacity of biosynthesis of histidine and a crucial role in metal homeostasis and virulence (Dietl et al., 2016); and the extra $NAD^+$ *de novo* biosynthesis pathway in *M. curtisii* could also enhance their virulence during host infection (Dom and Haesebrouck, 1992; Wang et al., 2019). On the contrary, two virulence-associated gene clusters, including genes associated with ABC-type guanosine uptake system NupNOPQ and lactate utilization, were only present in *M. mulieris*. Another noteworthy was the *nan* gene cluster for sialic acid catabolism (SAC). With these genes *M. mulieris* strains were more likely to consume host sialic acids as carbon source but could not cleave terminal Neu5Ac residues from host glycoconjugates (lacking *nanH*), whereas *M. curtisii* did just the opposite. This pattern perhaps implicated a cooperation between closely related species. However, this cooperation relationship seems not to be strictly necessary, as women with BV could harbor both or either of the two species (Holst, 1990). SAC associated genes have been detected in many BV-associated bacteria, which could enhance the pathogenicity of organisms by allowing easier invasion and destruction of tissues (Hardy et al., 2017; Jones, 2019; Li and Huang, 2022). These results could reinforce further discrimination of *Mobiluncus* species, perhaps by providing a simple and fast approach for identifying *M. curtisii* and *M. mulieris* using PCR or culture experiments, and in addition might facilitate the development of novel strategies to detect and prevent *Mobiluncus* infection of BV.

HGT is known to have great, perhaps the most conspicuous, impacts on bacterial diversity and speciation, especially for clinical microorganisms, where acquisition of foreign genes is crucial for pathogenicity (Smillie et al., 2011; Diard and Hardt, 2017; Arnold et al., 2021). In this study, we have used several methods to evaluate the HGT events. Firstly, both of the two *Mobiluncus* species have an open pangenome, which could be considered as an indicator of high HGT rates (Medini et al., 2005; Tettelin et al., 2008). Then, we reconstructed the evolutionary history of the genus. As expected, gene families undergoing gain events at ancestral nodes of species outnumbered those that experienced loss events. Therefore, differences in metabolism and pathogenicity between species have emerged. Finally, by using a BLAST-based HGT detection approach, we found that more than 5% of genes in each strain have suffered transfer events, and these genes perhaps further promoted the functional divergence between *Mobiluncus* species. Interestingly, no significant correlation between genome size and HGT frequency was observed, which probably means strains of *M. mulieris*, compared to *M. curtisii*, tend to acquire more dispensable genes, or meanwhile have suffered more gene loss events. Most of the transferred genes originated within the *Actinobacteria* phylum, with more from members of orders *Propionibacteriales*, *Corynebacteriales,* and *Micrococcales.* These orders have been reported to include many pathogenic species that could cause devastating diseases in humans and animals (Barka et al., 2016; Park et al., 2019), and also include microorganisms that are also present in the human vagina (Funke et al., 1997; Aleshkin et al., 2006; de Figueiredo Leite et al., 2010; Okoli et al., 2019). Taken together, these findings suggested that genome dynamic, mediated by gene gain and loss, might be an important strategy for *Mobiluncus* species diversification, host adaptation and pathogenicity.

Collectively, the present study largely extends the understanding of the genomic features, virulence and antibiotic resistance profiling, and evolution of the genus *Mobiluncus*. Our results also highlight the difference between *M. curtisii* and *M. mulieris*, providing more clues for distinguishing of the two species. Nevertheless, more experimental evidences are needed to verify these differences. Fully understanding the pathogenic potential of *Mobiluncus* strains remains a complex task with much to be explored in the future.

## DATA AVAILABILITY STATEMENT

The genomes analyzed in this study are all available in NCBI GenBank database with the accession numbers listed in **Supplementary Table 1**.

## AUTHOR CONTRIBUTIONS

YL designed the study, performed bioinformatic analyses, and wrote the draft manuscript. JL contributed to the conception of the study. YW and JL interpreted, discussed the results, and revised the manuscript. All authors contributed to manuscript revision and approved the submitted version.

## FUNDING

This study was supported by the National Natural Science Foundation of China (32100002).

## ACKNOWLEDGMENTS

We are grateful to Prof. Bin-Bin Xie (Shandong University) for his support to YL during funding application.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.939406/full#supplementary-material

## REFERENCES

Adams, M. D., Wagner, L. M., Graddis, T. J., Landick, R., Antonucci, T. K., Gibson, A. L., et al. (1990). Nucleotide sequence and genetic characterization reveal six essential genes for the LIV-I and LS transport systems of *Escherichia coli. J. Biol. Chem.* 265, 11436–11443. doi: 10.1016/S0021-9258(19)38417-0

Africa, C. W., Nel, J., and Stemmet, M. (2014). Anaerobes and bacterial vaginosis in pregnancy: virulence factors contributing to vaginal colonisation. *Int. J. Environ. Res. Public Health* 11, 6979–7000. doi: 10.3390/ijerph110706979

Agarwal, S., Gopal, K., Upadhyaya, T., and Dixit, A. (2007). Biochemical and functional characterization of UDP-galactose 4-epimerase from *Aeromonas hydrophila*. *Biochim. Biophys. Acta* 1774, 828–837. doi: 10.1016/j.bbapap.2007.04.007

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, D517–D525. doi: 10.1093/nar/gkz935

Aleshkin, V. A., Voropaeva, E. A., and Shenderov, B. A. (2006). Vaginal microbiota in healthy women and patients with bacterial vaginosis and nonspecific vaginitis. *Microb. Ecol. Health Dis.* 18, 71–74. doi: 10.1080/17482960600891473

Almeida, S., Sousa, C., Abreu, V., Diniz, C., Dorneles, E. M., Lage, A. P., et al. (2017). Exploration of nitrate reductase metabolic pathway in *Corynebacterium pseudotuberculosis*. *Int. J. Genomics* 2017:9481756. doi: 10.1155/2017/9481756

Andreae, C. A., Titball, R. W., and Butler, C. S. (2014). Influence of the molybdenum cofactor biosynthesis on anaerobic respiration, biofilm formation and motility in *Burkholderia thailandensis*. *Res. Microbiol.* 165, 41–49. doi: 10.1016/j.resmic.2013.10.009

Ankenbrand, M. J., Hohlfeld, S., Hackl, T., and Förster, F. (2017). AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput. Sci.* 3:e116. doi: 10.7717/peerj-cs.116

Arnold, B. J., Huang, I. T., and Hanage, W. P. (2021). Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* 20, 206–218. doi: 10.1038/s41579-021-00650-4

Arries, C., and Ferrieri, P. (2022). *Mobiluncus curtisii* bacteremia: case study and literature review. *Infect. Dis. Rep.* 14, 82–87. doi: 10.3390/idr14010009

Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H. P., et al. (2016). Taxonomy, physiology, and natural products of *Actinobacteria*. *Microbiol. Mol. Biol. Rev.* 80, 1–43. doi: 10.1128/MMBR.00019-15

Bizarro, C., and Schuck, D. (2007). Purine and pyrimidine metabolism in Mollicutes. *Genet. Mol. Biol.* 30, 190–201. doi: 10.1590/S1415-47572007000200005

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

Chai, Y., Kolter, R., and Losick, R. (2009). A widely conserved gene cluster required for lactate utilization in *Bacillus subtilis* and its involvement in biofilm formation. *J. Bacteriol.* 191, 2423–2430. doi: 10.1128/JB.01464-08

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

Cherpes, T. L., Hillier, S. L., Meyn, L. A., Busch, J. L., and Krohn, M. A. (2008). A delicate balance: risk factors for acquisition of bacterial vaginosis include sexual activity, absence of hydrogen peroxide-producing lactobacilli, black race, and positive herpes simplex virus type 2 serology. *Sex. Transm. Dis.* 35, 78–83. doi: 10.1097/OLQ.0b013e318156a5d0

Christensen, M., Borza, T., Dandanell, G., Gilles, A. M., Barzu, O., Kelln, R. A., et al. (2003). Regulation of expression of the 2-deoxy-D-ribose utilization regulon, *deoQKPX*, from *Salmonella enterica* serovar Typhimurium. *J. Bacteriol.* 185, 6042–6050. doi: 10.1128/JB.185.20.6042-6050.2003

Contreras-Moreira, B., and Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* 79, 7696–7701. doi: 10.1128/AEM.02411-13

Coudray, M. S., and Madhivanan, P. (2020). Bacterial vaginosis-A brief synopsis of the literature. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 245, 143–148. doi: 10.1016/j.ejogrb.2019.12.035

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, 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

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

Curtis, A. H. (1913). A motile curved anaerobic Bacillus in uterine discharges. *J. Infect. Dis.* 12, 165–169. doi: 10.1093/infdis/12.2.165

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. doi: 10.1371/journal.pone.0011147

Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78, 1–39. doi: 10.1128/MMBR.00035-13

de Figueiredo Leite, S. R., de Amorim, M. M., Calabria, W. B., de Figueiredo Leite, T. N., de Oliveira, V. S., Ferreira Junior, J. A., et al. (2010). Clinical and microbiological profile of women with bacterial vaginosis. *Rev. Bras. Ginecol. Obstet.* 32, 82–87. doi: 10.1590/s0100-72032010000200006

Dela Cruz, E. J., Fiedler, T. L., Liu, C., Munch, M. M., Kohler, C. M., Oot, A. R., et al. (2021). Genetic variation in Toll-like receptor 5 and colonization with flagellated bacterial vaginosis-associated bacteria. *Infect. Immun.* 89, e00060–20. doi: 10.1128/IAI.00060-20

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

Dietl, A. M., Amich, J., Leal, S., Beckmann, N., Binder, U., Beilhack, A., et al. (2016). Histidine biosynthesis plays a crucial role in metal homeostasis and virulence of *Aspergillus fumigatus*. *Virulence* 7, 465–476. doi: 10.1080/21505594.2016.1146848

Dietl, A. M., Binder, U., Bauer, I., Shadkchan, Y., Osherov, N., and Haas, H. (2020). Arginine auxotrophy affects siderophore biosynthesis and attenuates virulence of *Aspergillus fumigatus*. *Genes (Basel)* 11:423. doi: 10.3390/genes11040423

Dom, P., and Haesebrouck, F. (1992). Comparative virulence of NAD-dependent and NAD-independent *Actinobacillus pleuropneumoniae* strains. *J. Vet. Med. B.* 39, 303–306. doi: 10.1111/j.1439-0450.1992.tb01173.x

Funke, G., Hutson, R. A., Hilleringmann, M., Heizmann, W. R., and Collins, M. D. (1997). *Corynebacterium lipophiloflavum* sp. nov. isolated from a patient with bacterial vaginosis. *FEMS Microbiol. Lett.* 150, 219–224. doi: 10.1111/j.1574-6968.1997.tb10373.x

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

Gardner, H. L., and Dukes, C. D. (1955). *Haemophilus vaginalis* vaginitis: a newly defined specific infection previously classified non-specific vaginitis. *Am. J. Obstet. Gynecol.* 69, 962–976. doi: 10.1016/0002-9378(55)90095-8

Gatti, M., Aschbacher, R., Cimmino, C., and Valentini, R. (1997). Antigenic profiles for the differentiation of *Mobiluncus curtisii* and *Mobiluncus mulieris* by immunoblotting technique. *New Microbiol.* 20, 247–252.

Han, X., and Turnbough, C. L. Jr. (1998). Regulation of *carAB* expression in *Escherichia coli* occurs in part through UTP-sensitive reiterative transcription. *J. Bacteriol.* 180, 705–713. doi: 10.1128/JB.180.3.705-713.1998

Hardy, L., Jespers, V., Van den Bulck, M., Buyze, J., Mwambarangwe, L., Musengamana, V., et al. (2017). The presence of the putative *Gardnerella vaginalis* sialidase A gene in vaginal specimens is associated with bacterial vaginosis biofilm. *PLoS One* 12:e0172522. doi: 10.1371/journal.pone.0172522

Hatoum-Aslan, A., and Marraffini, L. A. (2014). Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. *Curr. Opin. Microbiol.* 17, 82–90. doi: 10.1016/j.mib.2013.12.001

Hillier, S. L. (1993). Diagnostic microbiology of bacterial vaginosis. *Am. J. Obstet. Gynecol.* 169(2 Pt 2), 455–459. doi: 10.1016/0002-9378(93)90340-o

Holst, E. (1990). Reservoir of four organisms associated with bacterial vaginosis suggests lack of sexual transmission. *J. Clin. Microbiol.* 28, 2035–2039. doi: 10.1128/jcm.28.9.2035-2039.1990

Hoyles, L., Collins, M. D., Falsen, E., Nikolaitchouk, N., and McCartney, A. L. (2004). Transfer of members of the genus *Falcivibrio* to the genus *Mobiluncus*, and emended description of the genus *Mobiluncus*. *Syst. Appl. Microbiol.* 27, 72–83. doi: 10.1078/0723-2020-00260

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

Javed, A., Parvaiz, F., and Manzoor, S. (2019). Bacterial vaginosis: an insight into the prevalence, alternative treatments regimen and it's associated resistance patterns. *Microb. Pathog.* 127, 21–30. doi: 10.1016/j.micpath.2018.11.046

Jiang, T., Gao, C., Ma, C., and Xu, P. (2014). Microbial lactate utilization: enzymes, pathogenesis, and regulation. *Trends Microbiol.* 22, 589–599. doi: 10.1016/j.tim.2014.05.008

Jones, A. (2019). Bacterial vaginosis: a review of treatment, recurrence, and disparities. *J. Nurse. Pract.* 15, 420–423. doi: 10.1016/j.nurpra.2019.03.010

Kenyon, C., Colebunders, R., and Crucitti, T. (2013). The global epidemiology of bacterial vaginosis: a systematic review. *Am. J. Obstet. Gynecol.* 209, 505–523. doi: 10.1016/j.ajog.2013.05.006

Lee, P. C., and Rietsch, A. (2015). Fueling type III secretion. *Trends Microbiol.* 23, 296–300. doi: 10.1016/j.tim.2015.01.012

Li, C. T., Liao, C. T., Du, S. C., Hsiao, Y. P., Lo, H. H., and Hsiao, Y. M. (2014). Functional characterization and transcriptional analysis of *galE* gene encoding a UDP-galactose 4-epimerase in *Xanthomonas campestris* pv. *campestris*. *Microbiol. Res.* 169, 441–452. doi: 10.1016/j.micres.2013.08.005

Li, Y., and Huang, Y. (2022). Distribution and evolutionary history of sialic acid catabolism in the phylum *Actinobacteria*. *Microbiol. Spectr.* 10:e0238021. doi: 10.1128/spectrum.02380-21

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 47, D687–D692. doi: 10.1093/nar/gky1080

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

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

Meltzer, M. C., Desmond, R. A., and Schwebke, J. R. (2008). Association of *Mobiluncus curtisii* with recurrence of bacterial vaginosis. *Sex. Transm. Dis.* 35, 611–613. doi: 10.1097/OLQ.0b013e318167b105

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Muzny, C. A., and Schwebke, J. R. (2016). Pathogenesis of bacterial vaginosis: discussion of current hypotheses. *J. Infect. Dis.* 214(Suppl. 1), S1–S5. doi: 10.1093/infdis/jiw121

Ngo, T. D., Perdu, C., Jneid, B., Ragno, M., Novion Ducassou, J., Kraut, A., et al. (2020). The PopN gate-keeper complex acts on the ATPase PscN to regulate the T3SS secretion switch from early to middle substrates in *Pseudomonas aeruginosa*. *J. Mol. Biol.* 432:166690. doi: 10.1016/j.jmb.2020.10.024

Nugent, R. P., Krohn, M. A., and Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J. Clin. Microbiol.* 29, 297–301. doi: 10.1128/jcm.29.2.297-301.1991

Okoli, A. C., Agbakoba, N. R., Ezeanya, C. C., Oguejiofor, C. B., and Anukam, K. C. (2019). Comparative abundance and functional biomarkers of the vaginal and gut microbiome of Nigerian women with bacterial vaginosis: a study with 16S rRNA metagenomics. *J. Med. Lab. Sci.* 29, 1–26.

Onderdonk, A. B., Delaney, M. L., and Fichorova, R. N. (2016). The human microbiome during bacterial vaginosis. *Clin. Microbiol. Rev.* 29, 223–238. doi: 10.1128/CMR.00075-15

Pandey, R., and Rodriguez, G. M. (2014). IdeR is required for iron homeostasis and virulence in *Mycobacterium tuberculosis*. *Mol. Microbiol.* 91, 98–109. doi: 10.1111/mmi.12441

Park, C. J., Smith, J. T., and Andam, C. P. (2019). "Horizontal gene transfer and genome evolution in the phylum *Actinobacteria*," in *Horizontal Gene Transfer: Breaking Borders Between Living Kingdoms*, eds T. G. Villa and M. Viñas (Cham: Springer International Publishing), 155–174.

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

Parte, A. C. (2018). LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. *Int. J. Syst. Evol. Microbiol.* 68, 1825–1829. doi: 10.1099/ijsem.0.002786

Peebles, K., Velloza, J., Balkus, J. E., McClelland, R. S., and Barnabas, R. V. (2019). High global burden and costs of bacterial vaginosis: a systematic review and meta-analysis. *Sex. Transm. Dis.* 46, 304–311. doi: 10.1097/OLQ.0000000000000972

Ramos, H. C., Rumbo, M., and Sirard, J. C. (2004). Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. *Trends Microbiol.* 12, 509–517. doi: 10.1016/j.tim.2004.09.002

Richter, M., and Rossello-Mora, 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

Roberts, M. C., Baron, E. J., Finegold, S. M., and Kenny, G. E. (1985). Antigenic distinctiveness of *Mobiluncus curtisii* and *Mobiluncus mulieris*. *J. Clin. Microbiol.* 21, 891–893. doi: 10.1128/jcm.21.6.891-893.1985

Roberts, M. C., Hillier, S. L., Schoenknecht, F. D., and Holmes, K. K. (1984). Nitrocellulose filter blots for species identification of *Mobiluncus curtisii* and *Mobiluncus mulieris*. *J. Clin. Microbiol.* 20, 826–827. doi: 10.1128/jcm.20.4.826-827.1984

Salinas, A. M., Osorio, V. G., Pacha-Herrera, D., Vivanco, J. S., Trueba, A. F., and Machado, A. (2020). Vaginal microbiota evaluation and prevalence of key pathogens in ecuadorian women: an epidemiologic analysis. *Sci. Rep.* 10:18358. doi: 10.1038/s41598-020-74655-z

Schwebke, J. R., and Desmond, R. (2005). Risk factors for bacterial vaginosis in women at high risk for sexually transmitted diseases. *Sex. Transm. Dis.* 32, 654–658. doi: 10.1097/01.olq.0000175396.10304.62

Schwebke, J. R., and Desmond, R. A. (2007). A randomized trial of the duration of therapy with metronidazole plus or minus azithromycin for treatment of symptomatic bacterial vaginosis. *Clin. Infect. Dis.* 44, 213–219. doi: 10.1086/509577

Schwebke, J. R., and Lawing, L. F. (2001). Prevalence of *Mobiluncus* spp among women with and without bacterial vaginosis as detected by polymerase chain reaction. *Sex. Transm. Dis.* 28, 195–199. doi: 10.1097/00007435-200104000-00002

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

Sha, B. E., Chen, H. Y., Wang, Q. J., Zariffard, M. R., Cohen, M. H., and Spear, G. T. (2005). Utility of amsel criteria, nugent score, and quantitative PCR for *Gardnerella vaginalis*, *Mycoplasma hominis*, and *Lactobacillus* spp. for diagnosis of bacterial vaginosis in human immunodeficiency virus-infected women. *J. Clin. Microbiol.* 43, 4607–4612. doi: 10.1128/JCM.43.9.4607-4612.2005

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34, D32–D36. doi: 10.1093/nar/gkj014

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244. doi: 10.1038/nature10571

Sobel, J. D. (2000). Bacterial vaginosis. *Annu. Rev. Med.* 51, 349–356. doi: 10.1146/annurev.med.51.1.349

Spiegel, C. A. (1987). Susceptibility of *Mobiluncus* species to 23 antimicrobial agents and 15 other compounds. *Antimicrob. Agents Chemother.* 31, 249–252. doi: 10.1128/AAC.31.2.249

Srinivasan, S., Morgan, M. T., Liu, C., Matsen, F. A., Hoffman, N. G., Fiedler, T. L., et al. (2013). More than meets the eye: associations of vaginal bacteria with gram stain morphotypes using molecular phylogenetic analysis. *PLoS One* 8:e78633. doi: 10.1371/journal.pone.0078633

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542. doi: 10.1093/bioinformatics/btl117

Taha, T. E., Hoover, D. R., Dallabetta, G. A., Kumwenda, N. I., Mtimavalye, L. A., Yang, L. P., et al. (1998). Bacterial vaginosis and disturbances of vaginal flora: association with increased acquisition of HIV. *AIDS* 12, 1699–1706. doi: 10.1097/00002030-199813000-00019

Taylor-Robinson, A. W., and Taylor-Robinson, D. (2002). Evaluation of liquid culture media to support growth of *Mobiluncus* species. *J. Med. Microbiol.* 51, 491–494. doi: 10.1099/0022-1317-51-6-491

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

Thorsen, P., Jensen, I. P., Jeune, B., Ebbesen, N., Arpi, M., Bremmelgaard, A., et al. (1998). Few microorganisms associated with bacterial vaginosis may constitute the pathologic core: a population-based microbiologic study among 3596 pregnant women. *Am. J. Obstet. Gynecol.* 178, 580–587. doi: 10.1016/s0002-9378(98)70442-9

Vinuesa, P., Ochoa-Sanchez, L. E., and Contreras-Moreira, B. (2018). GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus *Stenotrophomonas*. *Front. Microbiol.* 9:771. doi: 10.3389/fmicb.2018.00771

Wang, Q., Hassan, B. H., Lou, N., Merritt, J., and Feng, Y. (2019). Functional definition of NrtR, a remnant regulator of $NAD^+$ homeostasis in the zoonotic pathogen *Streptococcus suis*. *FASEB J.* 33, 6055–6068. doi: 10.1096/fj.201802179RR

Williams, M. J., Kana, B. D., and Mizrahi, V. (2011). Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. *J. Bacteriol.* 193, 98–106. doi: 10.1128/JB.00774-10

Wylensek, D., Hitch, T. C. A., Riedel, T., Afrizal, A., Kumar, N., Wortmann, E., et al. (2020). A collection of bacterial isolates from the pig intestine reveals functional and taxonomic diversity. *Nat. Commun.* 11:6389. doi: 10.1038/s41467-020-19929-w

Zeng, W., Ma, H., Fan, W., Yang, Y., Zhang, C., Arnaud Kombe Kombe, J., et al. (2020). Structure determination of CAMP factor of *Mobiluncus curtisii* and insights into structural dynamics. *Int. J. Biol. Macromol.* 150, 1027–1036. doi: 10.1016/j.ijbiomac.2019.10.107

Zhang, X., Bai, Y., Zhang, L., Draz, M. S., Ruan, Z., and Zhu, Y. (2020). Antimicrobial susceptibility and clonality of vaginally derived multidrug-resistant *Mobiluncus* isolates in China. *Antimicrob. Agents Chemother* 64, e00780–20. doi: 10.1128/AAC.00780-20

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., et al. (2014). PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* 30, 1297–1299. doi: 10.1093/bioinformatics/btu017

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

**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.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fmicb.2022.984582

# Identification of region of difference and H37Rv-related deletion in *Mycobacterium tuberculosis complex* by structural variant detection and genome assembly

Zhuochong Liu[1†], Zhonghua Jiang[1†], Wei Wu[1], Xinyi Xu[1], Yudong Ma[1], Xiaomei Guo[2], Senlin Zhang[2] and Qun Sun[1*]

[1]Key Laboratory of Bio-Resources and Eco-Environment of the Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China, [2]College of Biomass Science and Engineering, Sichuan University, Chengdu, China

*Mycobacterium tuberculosis* complex (MTBC), the main cause of TB in humans and animals, is an extreme example of genetic homogeneity, whereas it is still nevertheless separated into various lineages by numerous typing methods, which differ in phenotype, virulence, geographic distribution, and host preference. The large sequence polymorphism (LSP), incorporating region of difference (RD) and H37Rv-related deletion (RvD), is considered to be a powerful means of constructing phylogenetic relationships within MTBC. Although there have been many studies on LSP already, focusing on the distribution of RDs in MTBC and their impact on MTB phenotypes, a crumb of new lineages or sub-lineages have been excluded and RvDs have received less attention. We, therefore, sampled a dataset of 1,495 strains, containing 113 lineages from the laboratory collection, to screen for RDs and RvDs by structural variant detection and genome assembly, and examined the distribution of RvDs in MTBC, including RvD2, RvD5, and *cobF* region. Consistent with genealogical delineation by single nucleotide polymorphism (SNP), we identified 125 RDs and 5 RvDs at the species, lineage, or sub-lineage levels. The specificities of RDs and RvDs were further investigated in the remaining 10,218 strains, suggesting that most of them were highly specific to distinct phylogenetic groups, could be used as stable genetic markers in genotyping. More importantly, we identified 34 new lineage or evolutionary branch specific RDs and 2 RvDs, also demonstrated the distribution of known RDs and RvDs in MTBC. This study provides novel details about deletion events that have occurred in distinct phylogenetic groups and may help to understand the genealogical differentiation.

KEYWORDS

***Mycobacterium tuberculosis* complex, large sequence polymorphism, region of difference, H37Rv-related deletion, structural variant**

OPEN ACCESS

EDITED BY
Yufeng Wang,
University of Texas at San Antonio,
United States

REVIEWED BY
Prasit Palittapongarnpim,
Mahidol University, Thailand
Aboma Zewude,
United Arab Emirates University,
United Arab Emirates

*CORRESPONDENCE
Qun Sun
qunsun@scu.edu.cn

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

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 02 July 2022
ACCEPTED 15 August 2022
PUBLISHED 08 September 2022

CITATION
Liu Z, Jiang Z, Wu W, Xu X, Ma Y, Guo X, Zhang S and Sun Q (2022) Identification of region of difference and H37Rv-related deletion in *Mycobacterium tuberculosis complex* by structural variant detection and genome assembly. *Front. Microbiol.* 13:984582. doi: 10.3389/fmicb.2022.984582

## Introduction

*Mycobacterium tuberculosis* (MTB) and other members of the *Mycobacterium tuberculosis* complex (MTBC) are responsible for the development of tuberculosis (TB) in humans or animals. In addition to MTB, MTBC also comprises *M. africanum* (Vasconcellos et al., 2010), *M. canetti* (Soolingen et al., 1997), *M. orygis* (Pittius et al., 2012), *M. caprae* (Niemann et al., 2002), *M. bovis* (Garnier et al., 2003), *M. mungi* (Al, 2010), *M. suricattae* (Parsons et al., 2013), *M. microti* (Brodin et al., 2002), and *M. pinnipedii* (Cousins et al., 2003). With the exception of *M. canetti*, the remaining members of MTBC have selfsame 16S rRNA sequences and nearly identical genome sequences, and there is no horizontal gene transfer between MTBC strains, therefore MTBC is regarded as a conspicuous example of remarkable genetic homogeneity, and the extreme similarity proves that they share a common ancestor, but still differ in host preference, phenotype, and virulence (2003; Hirsh et al., 2004; Sanou et al., 2015). Among them, MTB and *M. africanum* are the most common causes of TB in humans, and prior studies have classified them into 9 primary lineages (lineage 1∼9), some of which are strongly geographically restricted (lineage 1, 3, and lineage 5∼9), while others (lineage 2, 4) are found globally distributed (Coscolla and Gagneux, 2014; Devis et al., 2020; Ngabonziza et al., 2020).

Several methods have been developed for molecular typing of MTBC strains, including IS6110 restriction fragment length polymorphism (IS6110 RFLP), mycobacterial interspersed repetitive units-variable number tandem repeats (MIRU-VNTR), and spacer region oligonucleotide typing (Spoligotyping). These methods have showed the high resolution and perform well in clinical testing, traceability, and re-infection detection. However, exorbitant diversity or, in some cases, excessive homogeneity has led to the application of these methods in phylogenetic analysis not entirely suitable for constructing reasonable phylogenetic relationships for MTBC. In addition to the molecular typing methods described above, Napier et al. (2020) provided a typing scheme based on single nucleotide polymorphism (SNP) consisting of 90 specific robust SNPs to cover 90 MTBC (sub-) lineages or species; although the SNP-defined lineages do not offer the same resolution as using the whole genome, they provide a valuable insight into the epidemiology of circulating strains.

Early comparative genomics studies used whole-genome microarrays or bacterial artificial chromosomes to identify areas with considerable diversity in the genomes of distinct MTBC lineages, known as large sequence polymorphism (LSP), which are frequently deleted in different strains (Brosch et al., 1998; Gordon et al., 1999). The associated deletion events are often unidirectional and irrecoverable, which will be inherited by all progeny strains. The LSPs are thus regarded as a valuable tool for establishing plausible phylogenetic relationships within MTBC, in addition to being utilized for molecular typing. LSPs incorporates region of RD and RvD, where RD refers to deletions relative to H37Rv and RvD refers to deletions in H37Rv relative to the rest of the strains.

Whole-genome sequencing (WGS) has achieved advancements in the study of MTB resistance, transmission kinetics, and phylogenetic analysis of MTBC as sequencing technology improves (Meehan et al., 2019). RD-Analyzer was the pioneer in the analysis of LSPs using WGS data to predict the species or lineage of MTBC strains based on 31 previously defined RDs (Kiatichai et al., 2016). Previous studies of RDs, as observed by Shitikov (Bespiatykh et al., 2021), have focused on discovering RDs within certain MTBC lineages, frequently without addressing the remaining members of the MTBC, or on studying the relationship of various members within the MTBC, although often only considering crucial genomic regions. Therefore, only a small number of RDs are available, and the distribution of RDs in MTBC lacks comprehensive understanding. Shitikov compiled a complete map of RDs in MTBC by collecting known RDs, examining the specificities of these RDs in a dataset encompassing several MTBC lineages, and identifying novel RDs. In addition, they developed RDscan, a pipeline for detecting RDs using WGS data.

In contrast with RDs, RvDs have received less attention using WGS data, possibly because of the discovery that RvDs frequently require genome assembly. Early studies identified that 5 RvDs (RvD1 ∼ RvD5) were present in *M. bovis* but absent in H37Rv, and it was suggested that the absence of RvD2 ∼ RvD5 could be due to recombination between adjacent isotropic IS6110 (Brosch et al., 1999; Gordon et al., 1999), increasing the difficulty of finding RvDs in assembled draft genomes because of the need to overcome the gap introduced by IS6110. Aside from the deletion of TbD1 in modern lineages and the retention of the *cobF* region in Lineage 8 and *M. canetti*, few studies have found that RvDs are linked to genealogical differentiation in MTBC (Brosch et al., 2002; Ngabonziza et al., 2020).

We sampled a dataset of 1,495 strains from the laboratory collection of MTBC strains with WGS data from 113 lineages, and RDs were screened using a pipeline consisting of multiple tools to detect deletions, while RvDs were screened by genome assembly. We identified 125 RDs and 5 RvDs, including 34 newly named RDs and 2 RvDs, specific to distinct phylogenetic groups. In addition, we discovered a complicated deletion in the RvD4496 region in the genome of lineage 5, where the RvD4496 was partially deleted and a 3.5 kb long fragment was absent in the downstream region. Most of the identified RDs and RvDs were highly specific to distinct phylogenetic groups and could be used as stable genetic markers in genotyping according to the results of specificity test in remaining strains. Further, we demonstrated the distribution of known RDs and RvDs in MTBC to show the details about deletion events, and this may help to understand the genealogical differentiation within MTBC.

# Materials and methods

## Dataset

WGS data for 11,713 MTBC strains were collected from the NCBI-SRA archive.[1] The SRA files were downloaded locally using sratoolkit (version: 2.11.3)[2] before being decompressed into pair-end FASTQ files. All WGS data obtained were quality controlled using Fastp (version: 0.23.1) (Chen et al., 2018). After mapping to reference genome, depth and sequencing coverage of WGS data were counted by bamdst,[3] and those with depth below 50 × and coverage below 90% were excluded. Then after SNP typing, a random sample of 1,495 strains from 113 lineages constituted the dataset used for screening RDs and RvDs. The work flow of the analysis process used for the sampled dataset and the remaining strains could be referred to **Supplementary material 9**.

## SNP calling and typing

A pipeline built in-house in the laboratory was applied to perform variant calling for all strains (reference genome H37Rv, NCBI accession number: GCF_000195955.2). The pipeline was as follows: quality-controlled WGS data were mapped to the reference genome using bwa-mem2 (version: 2.2.1) (Vasimuddin et al., 2019), before being sorted by samtools (version: 1.14) (Danecek et al., 2021) and PCR duplicates were marked using sambamba (version: 0.8.0) (Tarasov et al., 2015), respectively, followed by bamdst (version: 1.0.9) (see text footnote 3) to calculate the sequencing depth and coverage, and removed strains with below 90% coverage. SNP typing was performed using fast-lineage-caller (version: 0.3.2)[4] and a Python script based on the method of Clark (Napier et al., 2020). Only strains with consistent typing results in both methods were retained for subsequent analysis. Strains of lineage 1.1.1, lineage 2.2.1, lineage 5, and lineage 6 were classified into more specific sub-lineages according to the methods of Shitikov (Shitikov et al., 2017), Palittapongarnpim (Palittapongarnpim et al., 2018), and Coscolla (Devis et al., 2020). Labels in the NCBI-SRA archive of the strains belonging to animal-adapted lineages were reserved.

## Phylogenetic analysis

All SNP loci of all strains were sequentially aligned and each strain was filled at these loci with deletion substitutions of "–," while loci with deletion rates higher than 20% were removed to produce the final aligned fasta files. SNPs in PE/PPE family genes, known drug resistance genes, and non-SNP variation were removed. Possible tandem repeat regions in the H37Rv were identified by TRF (version: 4.09)[5] and SNPs located in these regions were removed. The phylogenetic tree was constructed by RAxML (version: 8.0.0) (Stamatakis, 2014) using a maximum likelihood method with 500 bootstraps, and finally visualized and modified using iTol.[6]

## Structural variant detection

Delly (version: 0.8.7) (Rausch et al., 2012), Manta (version: 1.6.0) (Chen et al., 2016) and SvABA (version: 1.1.3) (Wala et al., 2018) were used to detect deletions in strains with default parameters. Results of a single strain were merged by SURVIVOR (version: 1.0.7) (Jeffares et al., 2017), to combine deletions detected by different cnallers with breakpoints located within 200 bp of each other, and deletions greater than 200 bp by more than two callers were retained. Genome assembly was performed using Shovill (version: 1.1.0)[7] with default parameters, followed by SVIM-asm (version: 1.0.2) (Heller and Vingron, 2020) to detect structural variants and recorded deletions and insertions greater than 200 bp. Among the results, those associated with DR regions (direct repeat regions, cluster and regularly spaced CRISPR sequences for spolygotyping) were removed.

## Re-genotyping and structural variation filtering

All structural variants detected in individual strains were collected and merged by SURVIVOR to form a structural variant library and used for subsequent re-genotyping of all strains to reduce false negatives in individual strains. Possible differences in breakpoints between individual strains were temporarily ignored, and the exact breakpoints of these structural variants will then be confirmed manually with IGV (version: 2.11.2) (Robinson et al., 2011). Re-genotyping was performed for all strains using svtyper (version: 0.1.1) (Chiang et al., 2015). For deletions, the following steps were performed to filter the re-genotyping results to reduce false positives and to ensure that deletions occurred mainly in single copy regions. Specifically, bamdst was employed to count sequencing uncoverage for calculating proportion of uncoverage (ratio of the total length of the sequencing uncoverage within deletion to the total length of the deletion), and deletions were considered as true

[1] https://www.ncbi.nlm.nih.gov/sra/

[2] http://www.ncbi.nlm.nih.gov/books/NBK158900/

[3] https://github.com/shiquan/bamdst/

[4] https://github.com/farhat-lab/fast-lineage-caller/

[5] https://tandem.bu.edu/trf/trf409.linux64.download.html

[6] https://itol.embl.de/

[7] https://github.com/tseemann/shovill/

positives when the proportion of uncoverage was higher than 0.75. Deletions with overlapping range may interfere with the filtering, so the proportion of uncoverage within 200 bp flanking the deletions was examined to determine if extended deletions existed. Deletions were considered to be extended when the proportion of uncoverage within 200 bp flanking the deletion was higher than 25%. For insertions, insertion fragments caused by large fragment duplications were removed by sequence characterization to ensure that the insertion fragment was novel. Specifically, the specific sequences of insertions with relevant insertion sites (located within 200 bp from each other) were collected, and multiple sequences alignment were performed to confirm whether they were likely to be the same insertion and to obtain concordant sequences, followed by searching for similar sequence fragments in the reference genome using MMseqs2 (version: 13-14511).[8] The longest segment in "Query coverage" was found from the BLAST results with "Percent of Identity" higher than 75%, and the product of "Percent of identity" and "Query coverage" of this segment was taken as the total similar sequence proportion. When the proportion was inferior to 0.75, the insertion fragment was considered as a novel insertion. Then a reference sequence containing all novel insertions was constructed separately to determine their presence by checking the sequencing coverage. More precisely, same as checking whether deletions were true positive, we mapped the sequencing data to the reference sequence containing the novel insertions, and then calculated the sequencing coverage for each novel insertion by bamdst. When the proportion sequencing coverage was above 0.75, we assumed the presence of the novel insertion in the strain.

### Screening for region of differences and H37Rv-related deletions

The number of strains with certain deletions or insertions in each lineage was counted and the RDs and RvDs were screened based on the following criteria: when the number of strains in the lineage was less than 10, all strains should have the deletion, and when the number of strains was greater than 10, a maximum of 10% and no more than 5 strains were allowed to be free of the deletion (as these could be sequencing errors or false negatives). Later, we will examine the distribution of screened RDs and RvDs in MTBC to confirm whether they converge in specific lineages or phylogenetic clades.

### Breakpoint confirmation

A custom Python script was used to determine if both ends of the RDs were located in or near tandem repeat regions. For each RD and RvD, breakpoints were confirmed as follows: mapping the draft genome assembled by Shovill to H37Rv using minimap2 (version: 2.24) (Li, 2018), generating different SAM files by lineage, sorting and converting to BAM files using samtools. Then, IGV was used to visualize BAM files and the exact breakpoints of the deletion or insertion were set as those with the highest frequency. And for breakpoints located in repeat regions, the exact breakpoints were set as those with the largest deletion range.

### Structural variation annotation and covariance analysis

For RDs, genes included in deletions were identified by a custom Python script to confirm whether the deleted genes were essential according to DeJesus (DeJesus et al., 2017). For RvDs, covariance analysis was performed including H37Rv, lineage 8 (NCBI accession number: GCF_012923765.1), lineage 5 (NCBI accession number: GCF_905183075.1), and *M. canetti* (NCBI accession number: GCF_000253375.1). Sequences flanking the insertion site of RvDs were obtained by bedtools (version: 2.28.0) and then mapped to the rest of other reference genomes using minimap2 to generate SAM files. The SAM files were viewed through Pycharm, the best mappings were recorded as covariance regions, and the sequences of the covariance regions (insertion fragments and regions near the insertion sites) were obtained using bedtools. The genes contained within the covariance regions were identified by comparing the annotation files (GFF files) of each reference genome, before being aligned between the reference genomes by MMseqs2 to confirm the covariance of genes. Concordant sequences of RvDs reported by previous studies but not detected by Shovill and SVIM-asm were also acquired by covariance analysis, including RvD2, RvD5, and *cobF* regions (Brosch et al., 1999; Ngabonziza et al., 2020), and the reference genome of *M. bovis* (NCBI accession number: GCF_ 005156105.1) was additionally used here.

## Results

### Dataset and phylogenetic analysis

Variant calling and SNP typing were performed on the laboratory collection of 11,713 MTBC strains, and a dataset consisting of 1,495 strains from 113 lineages was randomly sampled (**Supplementary material 2**). There were less than ten strains in this dataset for 22 lineages, including several lineages with sub-lineages such as lineage 3.1, lineage 4.1, lineage 4.3, lineage 4.4, lineage 4.6, and lineage 4.6.1, but only one strain for *M. mungi*. The rest of the other lineages contained 11~31 strains. Since lineage 1.3 and its sub-lineage were not found in the laboratory collection, they were not included in the sampling dataset.

[8] https://github.com/soedinglab/MMseqs2

The 1,495 MTBC strains' phylogenetic relationships were inferred from 146,872 SNPs, with *M. canetti* as the root (**Figure 1**). The branch lengths of *M. canetti* were manually truncated to show the phylogenetic relationships and branch lengths of the remaining members in the MTBC. The recently discovered lineage 8 was thought to be separated from the remaining MTBC members before they diverged because of the presence of the *cobF* region (Ngabonziza et al., 2020). Phylogenetic analysis verified the phylogenetic status of lineage 8. Apart from lineage 8, the remaining MTBC members were divided into two main evolutionary branches, one for human-adapted lineages (lineage 1∼4 and lineage 7), and the other for the traditionally known *M. africanum* (lineage 5, 6, and lineage 9) and the animal-adapted lineages. A new lineage of *M. africanum*, lineage 9, was identified by Devis et al. (2020) as a sister lineage to lineage 6. Previous studies have split the animal-adapted lineages into four distinct evolutionary branches, A1 (*M. suricattae*, *M. mungi* as well as "chimpanzee" and "Dassie" bacillus, but "chimpanzee" bacillus was not included in this study), A2 (*M. microti* and *M. pinnipedii*), A3 (*M. orygis*) and A4 (*M. caprae* and *M. bovis*) (Brites et al., 2018). In the sampled dataset, the strains labeled as these species were easily distinguishable. Overall, phylogenetic analysis revealed that phylogenetic connections were mostly consistent with earlier research (Brites et al., 2018; Napier et al., 2020).

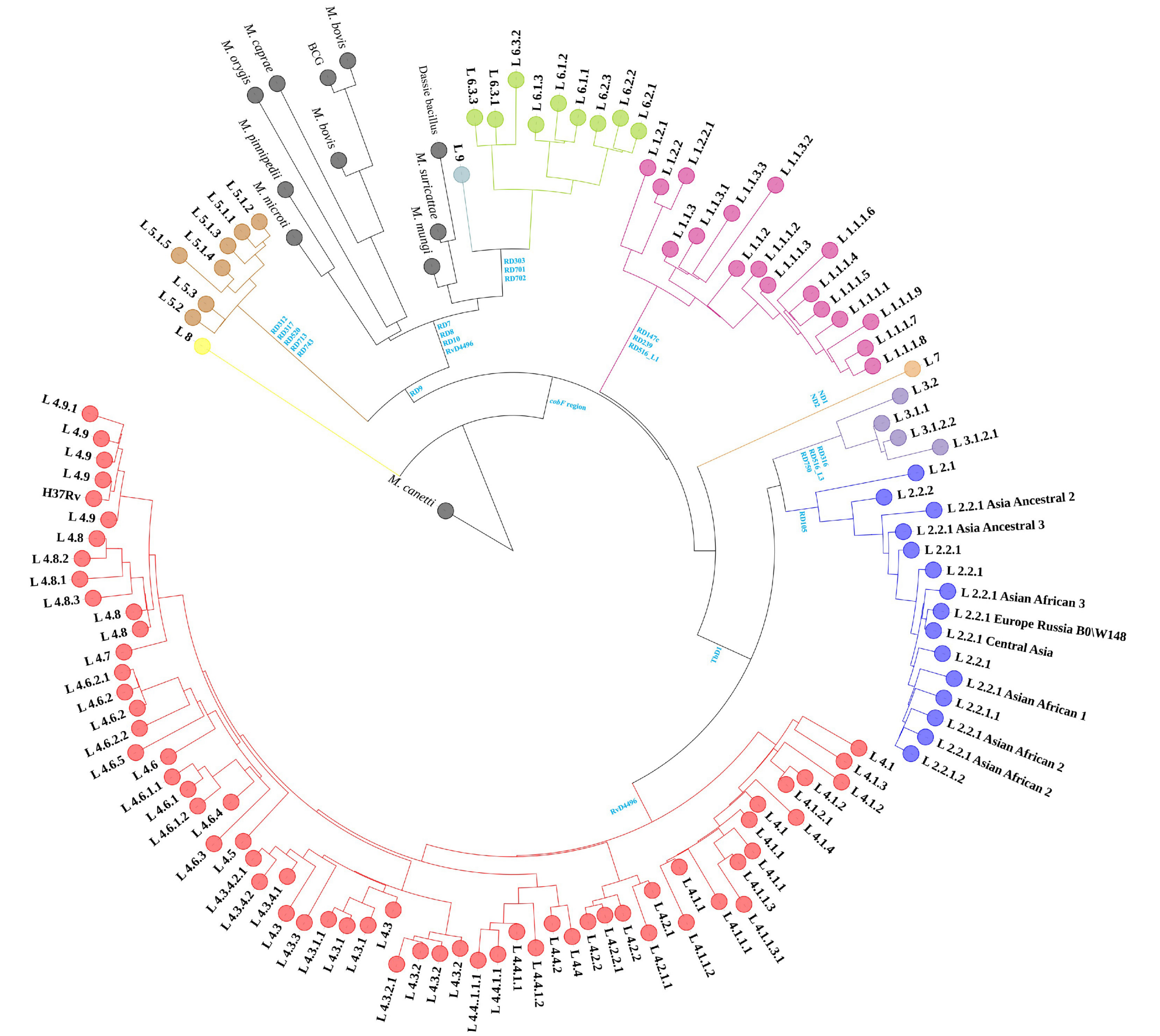


FIGURE 1
Maximum-likelihood phylogenetic tree of MTBC strains in sampled dataset. Clades were shrunk by lineage or sub-lineage, and the size of external nodes did not represent the number of strains.

## Structural variant detection, re-genotyping and filtering

Prior to re-genotyping using svtyper, the effectiveness of the structural variation detection pipeline, including the number of deletions detected in a single strain, the length of an individual deletion, the total length of detected deletions, the total length of sequencing uncoverage, and the detection efficiency of deletions (the detection efficiency of deletions is the ratio of the total length of detected deletions to the total length of sequencing uncoverage), was calculated (**Figure 2**). We also applied a filtering threshold of 0.75 for the proportion of sequencing uncoverage inside the deletions to guarantee that the discovered deletions were true positive.

An average of 5.26 (*SD* = 3.68) deletions per strain was observed, with the animal-adapted lineage having the greatest average number of deletions per strain (14.74, *SD* = 5.31) and lineage 4 having the lowest (2.67, *SD* = 1.84). Individual deletions were on average 3,073 bp long, with lineage 6 having the greatest (4,686 bp), followed by animal-adapted lineages (4,573 bp), and lineage 3 having the lowest (1,112 bp). The structural variation identification method discovered an average total length of 16,179 bp per strain, with lineage 3 having the least (5,960 bp) and the animal-adapted lineage having the highest (67,412 bp). For total sequencing uncoverage, bamdst detected an average of 54,293 bp per strain, with lineage 1 having the least (37,518 bp) and animal-adapted lineages having the highest (123,145 bp). The maximum of 118,985 bp of deletions in total length was detected in one *M. caprae* strain and the maximum of 437,036 bp of sequencing uncoverage in total length was detected in one lineage 5.1.4 strain. The deletion detection efficiencies ranged from 0 to 95.15%, and the average was only 31.43%.

## Lineage specific region of differences and H37Rv-related deletions

A total of 125 RDs and 5 RvDs were screened for specific lineages or evolutionary branches (**Figure 3**), with 91 RDs and 3 RvDs already reported by previous studies (**Supplementary material 3**). The 34 newly identified RDs belonging to MTBC members other than *M. canetti* and lineage 8 were named, while the 2 newly defined RvDs were named RvD533 and RvD4496 based on their length. In all, 54 RDs belonged to the evolutionary branch of *M. africanum* and animal-adapted lineages, while the remaining 70 RDs correspond to *M. tuberculosis.* A total of 17 newly designated RDs MTBC were determined to be specific to 17 lineages for which no RDs had previously been found, such as lineage 4.2.1.1, lineage 4.6.3, lineage 1.2.2.1, and lineage 6.2.3.

According to its discoverer, the recently discovered genealogy, lineage 9, contains a deleted region spanning from Rv1762c to Rv1765, since Rv1763 and Rv1764 are presumptive IS6110-4 in H37Rv (Devis et al., 2020). In the region upstream of Rv1762c, Rv1755 ∼ Rv1757 are putative IS6110-3. Lineage 2, a sub-lineage of lineage 4, lineage 5, and *M. orygis*, had a high frequency of deletions in the region between and downstream of the two IS6110s, whereas *M. orygis* had a distinct range of deletions (**Supplementary material 4—Figure 1**). Previous studies have noted deletions in this region including RD152 (Tsolaki et al., 2004) and RD14 (Gordon et al., 1999), however, no deletions in this region have been identified by the structural variation detection pipeline in our study. We discovered that lineage 9 and its sister evolutionary branch, lineage 6, shared the identical pattern of RDs, with no RDs unique to each. However, the independent absence of RD11 (prophage phiRv2) in several lineages might be used to distinguish between the two, with all lineage 6 strains deleting RD11, while this deletion was detected in one lineage 9 strain (four in total) and additional Lineage 9 strains were needed to be confirmed whether the deletion of RD11 was common. We were only able to detect the specificities of these RDs belonging to lineage 8 since there were only two strains of lineage 8 in the dataset. We detected 5 of the 8 non-IS6110 deletions identified by the discoverer of lineage 8, with the exception of RD3 as well as the RD14 region, all of which differed from the previously identified range. In addition, we identified 5 additional RDs, but their authenticity and specificity need to be further determined.

Of the 5 RvDs, RvD1, TbD1, and the *cobF* region have been confirmed by previous studies (Gordon et al., 1999; Brosch et al., 2002; Ngabonziza et al., 2020). RvD1 was confirmed to be lacking in lineage 4.8 and lineage 4.9 in our study by checking its distribution in MTBC. Consistent with previous studies, TbD1 was absent in strains of the modern lineage (lineage 2∼4) and the *cobF* region was absent in all strains except *M. canetti* and lineage 8. RvD533 was deleted in lineage 4.7, lineage 4.8, and lineage 4.9. RvD4496 was absent in lineage 4, lineage 6, lineage 9, and animal-adapted lineages. Two other RvDs, RvD2 and RvD5, missing due to IS6110 recombination (Brosch et al., 1999), were examined for their distribution in MTBC (**Supplementary material 3**) (the congruent sequences of RvD2 and RvD5 in different genomes please see **Supplementary material 5**). RvD2 was shown to be deleted independently in multiple lineages, with some lineages having all strains lacking RvD2 and others having only some strains deleted RvD2, demonstrating that RvD2 is not a suitable lineage-specific RvD. In contrast to RvD2, RvD5 was only missing in partial strains within lineage 4.9, which is the closest lineage to H37Rv. In a previous study, in another reference genome candidate, H37Ra, also belonging to lineage 4.9, RvD5 was not missing (Brosch et al., 1999).

## Different types of deletions

Depending on the nature of the deletion and the presence or absence of repeat sequence or mobile elements at either end,

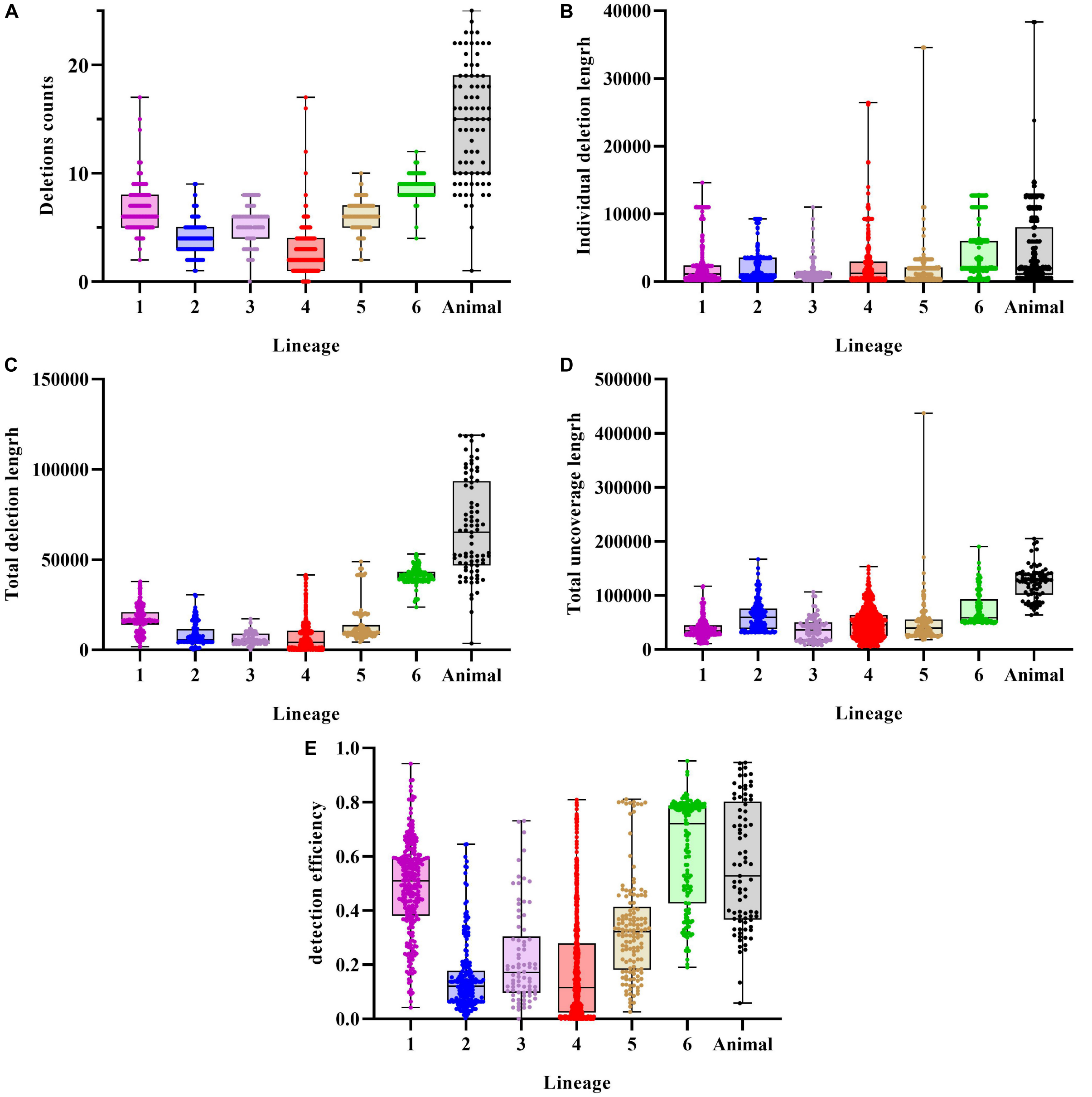


FIGURE 2
Characteristics of deletions in MTBC strains. **(A)** Deletions per genome distribution among MTBC strains. **(B)** Individual deletion length distribution among lineages. **(C)** Total deletions length per genome distribution among MTBC strains. **(D)** Total uncoverage length per genome distribution among MTBC strains. **(E)** Detection efficiency of deletions distribution among MTBC strains.

deletions could be divided into three categories: deletions of mobile elements, for example, prophages and IS6110; deletions with mobile elements or repeat sequence at either end and those without repeat sequence or mobile elements at either end. The first two represent unstable regions of the genome.

The 2 prophages in H37Rv, phiRv1, and phiRv2, referred to as RD3 and RD11, respectively, in the RD system, were thought to be independently absent in different lineages (Fan et al., 2016). The structural variation detection pipeline we adopted allowed us to detect the deletion of RD3 and RD11 and to obtain precise breakpoints. In total, RD3 was deleted in 905 strains from 97 lineages. While the deletion of RD11 was detected in all lineage 1.1, lineage 6, lineage 7, animal-adapted A3, and A4 evolutionary branches. Outside of these lineages, the deletion was only detected in 41 samples from 13 lineages. Therefore, RD11 is still considered a stable RD for the above-mentioned lineages. IS6110 is a multi-copy mobile element in the genome and deletions of IS6110 at a single locus does not result in sequencing uncoverage. So that, in our study, deletions or insertions of IS6110 itself were ignored.

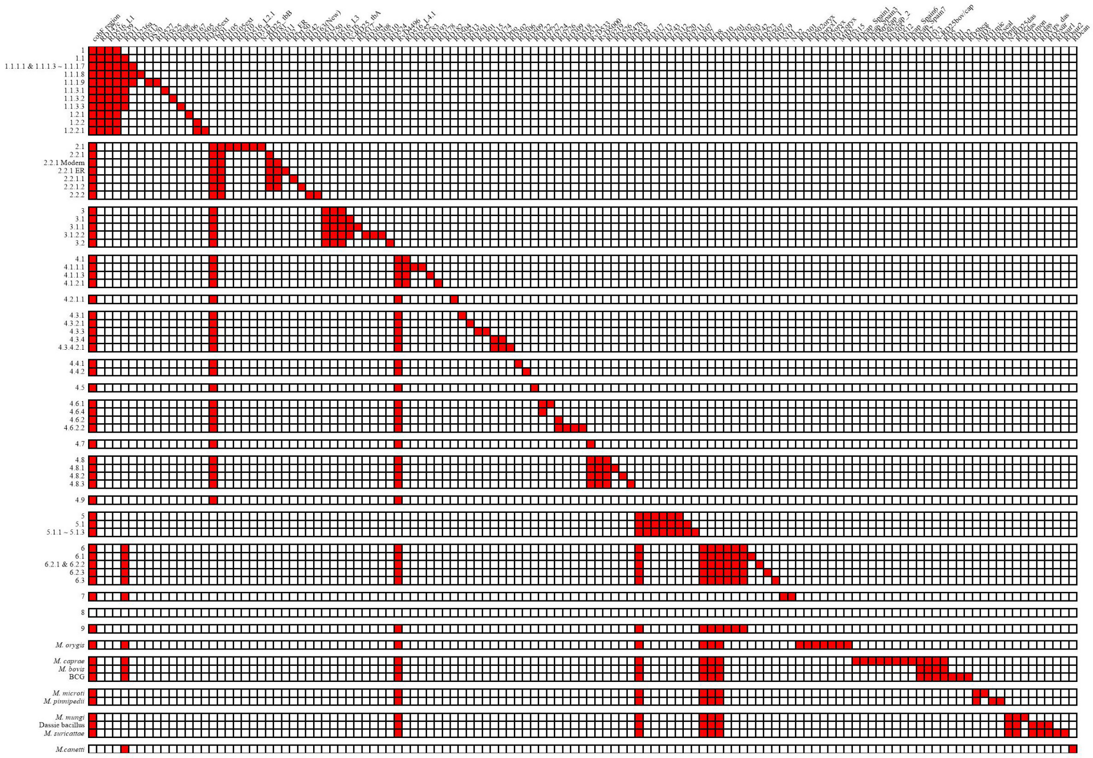


FIGURE 3
Deletion patterns of RDs and RvDs within MTBC.

In addition to RD3 and RD11, by comparing to the result of TRF and the annotation of reference genome, we analyzed the presence of repeat sequence and mobile elements at either end of the remaining deletions to distinguish between the latter two deletion types. Only 8 breakpoints in 7 RDs were found in repeat sequence or mobile elements, while repeat sequence or mobile elements were present within 200 bp at either end of another 23 RDs, including deletions in the RD5 region and RD1 region. Deletions in the RD5 region were of more concern in this category, which was thought to result in reduced virulence in humans, and inconsistent ranges have been observed in different animal-adapted lineages (Ates et al., 2018; **Supplementary material 4—Figure 2**). The remaining 94 RDs did not have any repeat sequence or mobile elements at either end.

With the exception of RvD2 and RvD5, the five RvDs mentioned above had no tandem repeat sequences or mobile elements near the insertion sites. Although we could not find an exact insertion site of the *cobF* region, and we checked its probable insertion region (105.3 kb).

## Deletions with overlapping range

19 groups of 57 RDs with overlapping ranges were identified (**Figure 4**), some of which may be associated with lineage differentiation. We could assume that different lineages, in which overlapping ranges of deletions occurred, if they could be divided into monophyletic groups in the phylogenetic tree, suggest that deletions of overlapping range may have occurred in a common ancestor, followed by different deletion extensions during subsequent divergence. For example, RD105 and RD105ext, classical RDs of lineages 2.2 and 2.1, respectively, with RD105ext having an extended deletion range to both sides compared to RD105. It is possible that deletion of RD105 was taken place in the common ancestor of lineage 2, and the extension occurred to form RD105ext during the differentiation into lineage 2.1 (**Figure 4A**). The same may also take place in RD505 and RD505ext (**Figure 4B**), RD7 and RD713 (**Figure 4D**), the deletions in RD5 region (**Supplementary material 4—Figure 1**) and the deletions in RD1 region (**Supplementary material 6—i**). RD505 and RD505ext are specific to lineage 1.2.2 and its sub-lineage 1.2.2.1, respectively, and their positions at the 3′-terminus are identical, but the deletion extended to form RD505ext in the further differentiated lineage 1.2.2.1. RD7 and RD713 are specific to two evolutionary branches, the evolutionary branch consisting of lineage 6, lineage 9, and animal-adapted lineages, and the evolutionary branch containing lineage 5, respectively. However, only a limited range of overlap existed between the two. In BCG, deletion extended at one end of

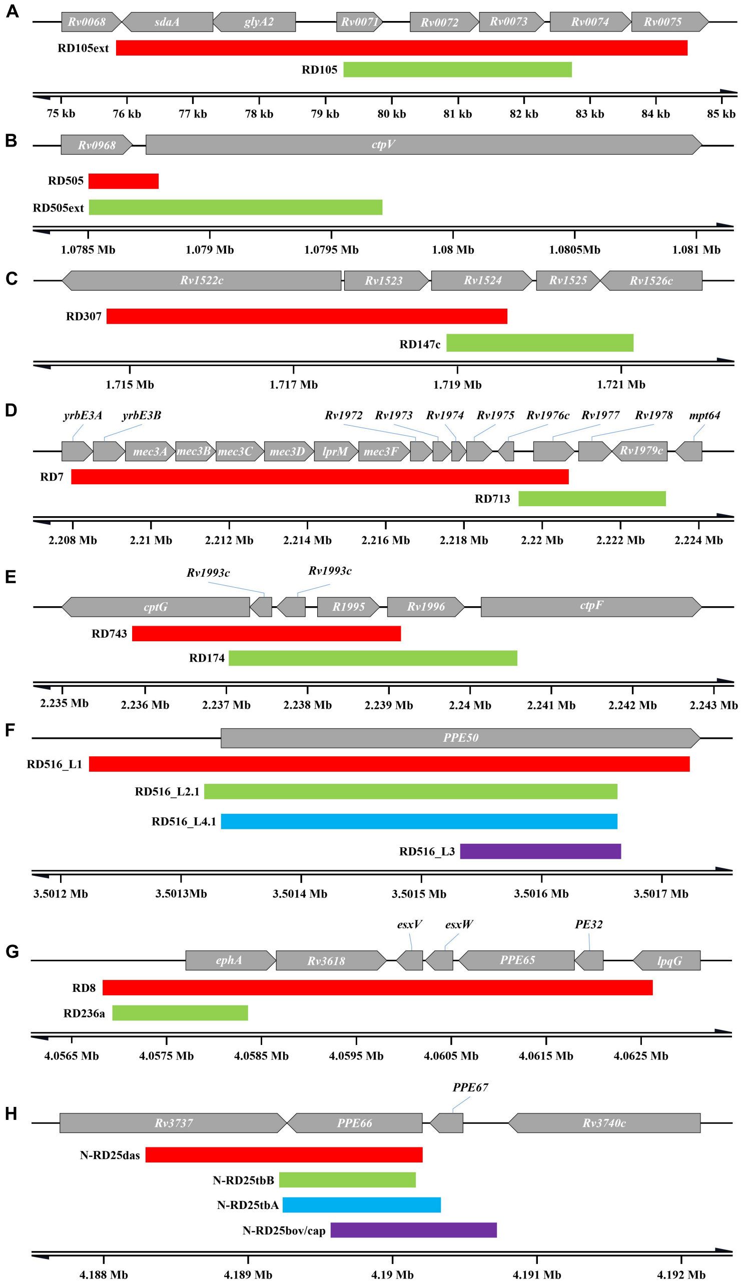


FIGURE 4

Overlapping RDs. Only the RDs mentioned in the text are shown, and do not include RD317 and RD306, and the deletion in RD5 region and RD1 region. **(A)** RD105 and RD105ext. **(B)** RD505 and RD505ext. **(C)** RD307 and RD147c. **(D)** RD7 and RD713. **(E)** RD 743 and RD174. **(F)** RDs in RD516 region. **(G)** RD8 and RD236a. **(H)** RDs in N-RD25 region.

RD5, while RD5 its self-corresponds to *M. bovis* and *M. caprae*. RD5oryx corresponds to *M. orygis*, however, the position of the 5′-terminus appears to change significantly amongst strains. Among the A1 evolutionary branch of "Dassie" bacillus, *M. mungi* and *M. suricattae*, deletions of the RD5 region appear to be more complicated. RD5das corresponds to *M. mungi* and was found in 2 of the 3 "Dassie" bacillus strains, while RD5sur, with a longer deletion range, corresponds to *M. suricattae* but was also found in 1 "Dassie" bacillus strains. The last group is the RD1 region with deletions RD1mon and RD1das. RD1mon corresponds to *M. mungi*, while RD1das corresponds to "Dassie" bacillus and *M. suricattae*. Although both BCG and *M. microti* have deletions in RD1 region, it was clear that they belong to different evolutionary branches from "Dassie" bacillus, *M. mungi* and *M. suricattae*, and did not constitute a monophyletic group.

The remaining RDs with overlapping ranges (**Supplementary material 6**) were identified in significant different evolutionary branches. For example, RD317 (lineage 5) and RD306 (lineage 4.4.1), RD307 (lineage 5.1.1 ∼ 5.1.3) and RD147c (lineage 1) (**Figure 4C**), RD743 (lineage 5) and RD174 (lineage 4.3.4) (**Figure 4E**), RD8 (lineage 6, lineage 9 and animal-adapted lineages) and RD236a (sub-lineages of lineage 1.1.1 except lineage 1.1.1.2) (**Figure 4G**). The N-RD25 (**Figure 4H**) area and the PPE50 gene (**Figure 4F**) are two other sites where deletions are linked to various lineages. N-RD25tbA corresponds to lineage 3; N-RD25tbB corresponds to lineage 2.1; N-RD25bovis/caprae corresponds to *M. bovis* and *M. caprae*, and N-RD25das corresponds to animal-adapted A1 evolutionary branch. There are four deletions in the PPE50 gene. RD516-L1 corresponds to lineage 1; RD516-L2.1 corresponds to lineage 2.1; RD516-L3 corresponds to lineage 3, and RD516-L4.1 corresponds to lineage 4.1.

For RvDs, when performing covariance analysis (**Supplementary material 7**), we identified a complicated deletion in RvD4496 region in the genome of lineage 5 (**Figure 5** and **Supplementary material 3**—**Table 1**). RvD4496 is partially deleted (the second half of the 5′-terminus), with deletion extending downstream to a 3.5 kb fragment in H37Rv.

## The importance of missing genes

The importance of the genes affected by RDs was confirmed in comparison to the study of DeJesus, although only for *in vitro* culture (**Supplementary material 8**). Only two genes (*Rv2017* and *Rv3902c*) were identified as essential among the 324 genes partly or totally eliminated owing to RDs (except RD528). *Rv2017* is presumed to be a transcriptional regulator. The function of *Rv3902c* is unknown, but studies suggest that it may be associated with virulence by co-expression with *EsxF*, *EsxE,* and *Rv3903c* (Danilchanka et al., 2014).

## Examining the specificity of region of differences and H37Rv-related deletions

We examined the specificities of these deletions in the remaining laboratory collection of 10,218 strains by assessing sequencing coverage (**Supplementary material 3**). 41 RDs were absent in strains outside of the specific lineages (excluding RD11, and deletions occurring in strains only classified into the upper lineage are also not counted). Since we were not concerned with the breakpoints in the genome of these strains, it could not be determined whether the exact range of these deletions was consistent with the RDs in range. Deletion of the RD5 region was detected in up to 168 strains of non-animal-adapted lineages; the next most common deletion was RD701, which was detected in 81 strains outside of lineage 6 and lineage 9; the remaining RDs were detected in no more than 35 strains outside of the non-specific lineage. Some of the RDs linked with the animal-adapted A3 and A4 evolutionary branches, such as those specific to *M. orygis* and *M. caprae*, respectively, were missing in some strains labeled as *M. bovis*, which might be due to labeling mistakes. *M. caprae* and *M. orygis* are closely related to *M. bovis*, hence in the early studies, they were often referred to as *M. bovis* until they were clearly identified. For the 5 RvDs, deletions were detected among non-specific lineages, except for *cobF* region.

In addition, some of the RDs might not be at the same node as the specific SNPs used for SNP typing, as evidenced by the detection of deletions in strains classified as upper lineage to the specific lineage, or the presence of true-negative non-deleted strains in the specific lineage. Further phylogenetic analysis of these strains is required to determine whether the RDs correspond to the specific SNPs. The RDs associated with *M. caprae* were the most prominent, and we performed a phylogenetic analysis of the laboratory collection of 32 *M. caprae* strains (**Figure 6**) in the same way as we did for other strains, confirming the occurrence of these RDs follows a pattern.

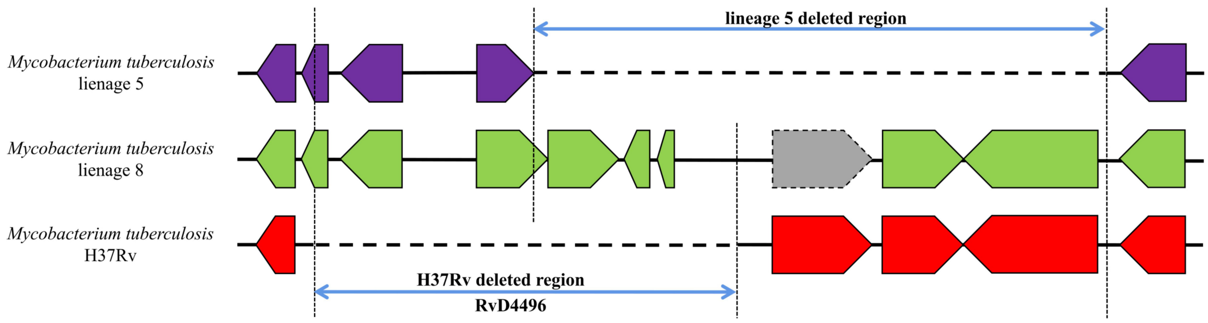

lineage 5 deleted region
Mycobacterium tuberculosis lienage 5
Mycobacterium tuberculosis lienage 8
Mycobacterium tuberculosis H37Rv
H37Rv deleted region
RvD4496

FIGURE 5
Gene covariation of RvD4496 region in lineage 5, lineage 8 and H37Rv. Gray block represents pseudogenes.

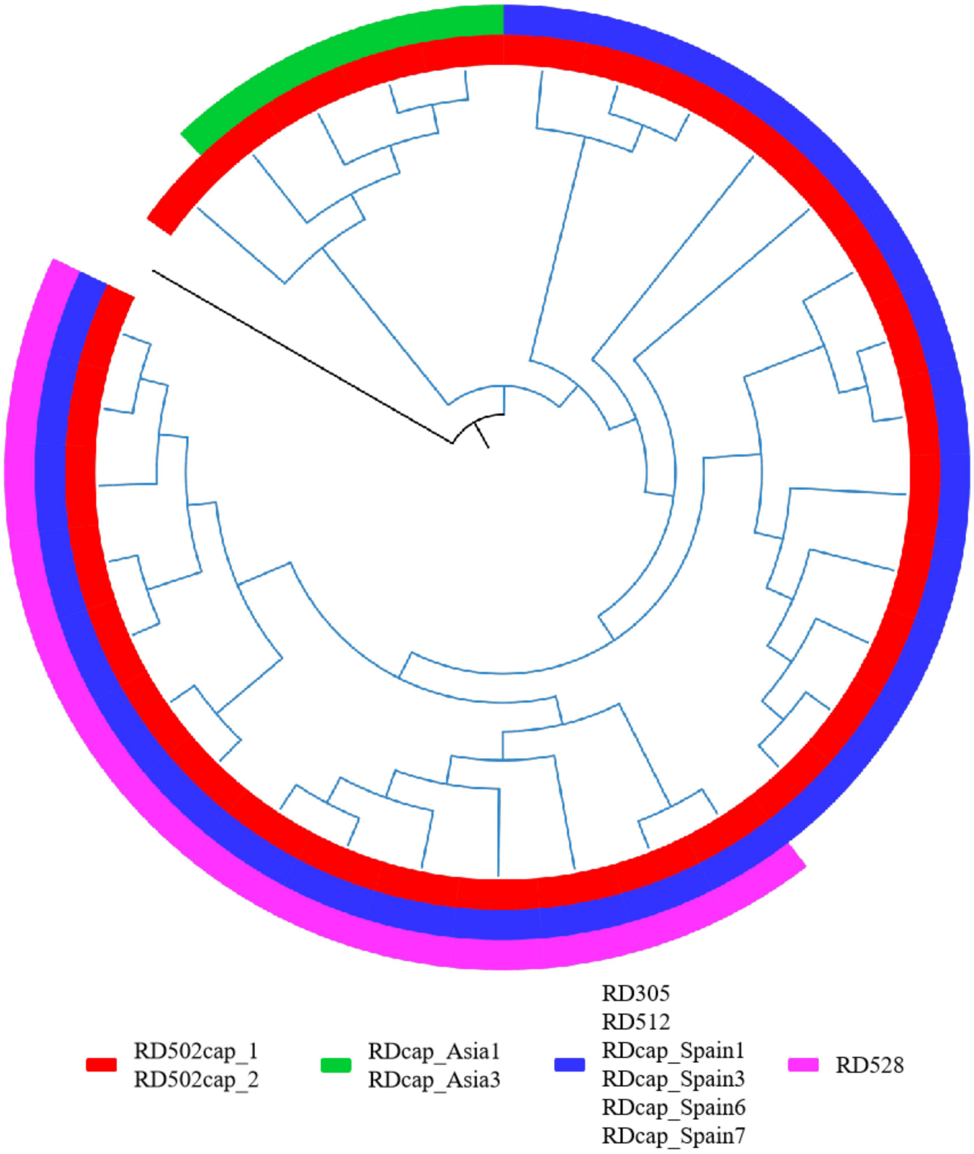


FIGURE 6
RDs in *M. caprae*. The blue branch represents the *M. caprae* strains and the black is *M. canetti* as the root. RD502cap_1 and RD502cap_2 are missing in all strains and the remaining RDs are only deleted in partial strains. RD528 is the largest deletion in length (38,328 bp) detected in this study and is deleted in 10 strains within the sampled dataset. Moreover, deletion of RD528 is detected in 4 additional strains out of all *M. caprae* strains.

## Discussion

Despite the use of several detection modalities, the results demonstrated a poor detection efficiency, which might be attributed to the intrinsically short read length of next-generation sequencing. Multiple algorithms for structural variant detection using next-generation sequencing data have a high proportion of false detection due to errors in base calling, alignment, or *de novo* assembly, especially in repetitive sequences that cannot be spanned by short sequencing reads. Long read lengths obtained by single-molecule sequencing techniques have recently been employed to discover structural variants in human samples to overcome the disadvantages (Chaisson et al., 2015; Pendleton et al., 2015). However, the high cost and low throughput of this method now prevent it from being widely used.

In contrast to phiRv2 (RD11), phiRv1 (RD3) is nevertheless of interest despite the fact that it cannot be considered a stable RD. It is thought to be incapable of encoding infectious phage particles (Fan et al., 2016), but may still have an active integration/excision system that enables transposition in the genome, and we found suspicious transposition-like phenomena in some strains, similar to IS6110 with multiple copies in the genome. We discovered breakpoints at the original position of RD3 using structural variation detection, indicating its deletion, although sequencing coverage revealed that it was not absent. Correspondingly, we found breakpoints in these strains indicated the translocation of RD3 (insertion sites 103.7 and 388.4 kb) which was comparable with previously observed RD3 insertion sites (Bibb and Hatfull, 2002).

A mobile element such as IS6110 is one of the causes of the genome assembly gaps. The IS6110 insertion found in TbD1 in the genome of lineage 8, prevented us from detecting the whole TbD1 in the assembled draft genome. And recombination between adjacent IS6110s might result in deletions of regions within, and such deletions were difficult to be detected due to the difficulty of identifying the accurate breakpoints. The presence of mobile elements or repeat sequences at either end of deletions represented the unstable region in MTBC genome, and deletions in these regions might occur independently in strains of different lineages and had inconsistent deletion ranges. This may explain some of the overlapping ranges of RDs that may result in some RDs being reported as missing in strains from non-specific lineages.

Although we have found RDs or RvDs accordingly in most of the evolutionary branches, it is still difficult to determine whether these deletions played a key driving role in the generation of these lineages at this stage. Deletion of the TbD1 increased the resistance of MTB to oxidative stress and hypoxia and enhanced its survival in granulomas, which is one of the primary drivers of modern lineages globally widespread (lineage 2∼4) (Bottai et al., 2020). We assume that the absence of some RDs and RvDs may have given MTB strains a competitive advantage in transmission, and pathogenicity, leading to the formation of new lineages. Alternatively, the partial range overlap of RDs that occurs independently in strains of distinct lineages might be the consequence of convergent evolution due to the same selection pressure. It's unclear why distinct deletions of the RvD4496 occurred in various evolutionary branches, but its intricate deletion in lineage 5 demonstrates limits beyond those of a single reference genome. The structural variant identification software Giraffe, developed by Sirén et al. (2021) provided good insights into the ability to map sequencing data to multiple reference genomes simultaneously to obtain more diverse and accurate genotyping results. The simultaneous use of reference genomes from multiple different lineages, including H37Rv, may help identify more structural variants associated with MTBC lineage differentiation.

Moreover, in comparative genomics investigations with NTM, *M. canetti* had earlier found the deletion of the *cobF* region in the MTB genome (2013) (Supply et al., 2013), which might result in MTBC's inability to synthesize vitamin B12 like other mycobacteria, which could be taken as a foreshadowing of lineage 8 discovery (2020). It can be assumed that in lineages with multiple RDs, the deletion events do not occur simultaneously and therefore intermediates may still be present. In the evolutionary branch with deletion RD7-RD8-R10, hosts jump between humans and animals may have occurred during evolution repeatedly, and the discovery of intermediates may assist in understanding the evolutionary history and the interaction of MTB with the host (Brites and Gagneux, 2015).

## Conclusion

We implemented multiple methods to search structural variations in the MTBC genomes and identified 125 RDs and 5 RvDs, including 34 newly identified RDs (RD501∼RD527 and RD2_ER while RD502 and RD516 contain multiple deletions) and 2 RvDs (RvD533 and RvD4496), specific to distinct phylogenetic groups. Thereinto, unreported RDs and RvDs were discovered in several new lineages and recognized sub-lineages. Further, we examined the distribution of RDs and RvDs in MTBC, and the results suggested that most of RDs and RvDs are persistent traits in parts of the MTBC lineage. Analysis of the RvD4496 region for lineage 5 revealed a complicated deletion, demonstrating the limitations of using only a single reference genome in comparative genomics research. The distribution of partial RDs with overlapping ranges in the phylogenetic tree implied that convergent evolution might result in the absence of identical regions in the genome due to exposure to the same selection pressure. Furthermore, as this study was performed *in silico* and the results need to be validated experimentally and evaluated using a dataset with additional 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 in the article/**Supplementary material**.

## Author contributions

QS and ZL designed the research work. ZL and ZJ performed the research activities and wrote the manuscript. ZL, ZJ, and WW analyzed the data and validated. QS, WW, XX, YM, and XG edited the manuscript submitted. All authors have given their approval for publication.

## Funding

This work was supported by the National Key Research and Development Projects (2019YFE0103800 and 2020YFB0704502) and the Science and Technology

Program of Sichuan Province (2021YFH0060 and 2022YFQ0100).

## Acknowledgments

We would like to thank Zhen Liu and Ling Hu for their linguistic assistance during the preparation of this 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.

## 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.2022.984582/full#supplementary-material

## References

Al, K. (2010). Novel *Mycobacterium tuberculosis* complex pathogen. *M. mungi. Emerg. Infect. Dis* 16, 1296–1299. doi: 10.3201/eid1608.100314

Ates, L. S., Sayes, F., Frigui, W., Ummels, R., Damen, M. P., Bottai, D., et al. (2018). RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. *PLoS Pathog.* 14:e1007139. doi: 10.1371/journal.ppat.1007139

Bespiatykh, D., Bespyatykh, J., Mokrousov, I., and Shitikov, E. (2021). A Comprehensive Map of *Mycobacterium tuberculosis* Complex Regions of Difference. *mSphere* 6:e0535-21. doi: 10.1128/mSphere.00535-21

Bibb, L. A., and Hatfull, G. F. (2002). Integration and excision of the *Mycobacterium tuberculosis* prophage-like element, φRv1. *Mol. Microbiol* 45, 1515–1526. doi: 10.1046/j.1365-2958.2002.03130.x

Bottai, D., Frigui, W., Sayes, F., Di Luca, M., Spadoni, D., Pawlik, A., et al. (2020). TbD1 deletion as a driver of the evolutionary success of modern epidemic *Mycobacterium tuberculosis* lineages. *Nat. Commun.* 11, 1–14. doi: 10.1038/s41467-020-14508-5

Brites, D., and Gagneux, S. (2015). Co-evolution of *Mycobacterium tuberculosis* and Homo sapiens. *Immunol. Rev.* 264, 6–24. doi: 10.1111/imr.12264

Brites, D., Loiseau, C., Menardo, F., Borrell, S., Boniotti, M. B., Warren, R., et al. (2018). A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* complex. *Front. Microbiol.* 9:2820. doi: 10.3389/fmicb.2018.02820

Brodin, P., Eiglmeier, K., Marmiesse, M., Billault, A., and Brosch, R. (2002). Bacterial artificial chromosome-based comparative genomic analysis identifies *Mycobacterium microti* as a natural ESAT-6 deletion mutant. *Infect. Immun.* 70, 5568–5578. doi: 10.1128/IAI.70.10.5568-5578.2002

Brosch, R., Gordon, S. V., Billault, A., Garnier, T., and Cole, S. T. (1998). Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect. Immun.* 66, 2221–2229. doi: 10.1128/IAI.66.5.2221-2229.1998

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., et al. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci.* 99, 3684–3689. doi: 10.1073/pnas.052548299

Brosch, R., Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, S. T., and Gordon, S. V. (1999). Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated M. tuberculosis H37Ra strain. *Infect. Immun.* 67, 5768–5774. doi: 10.1128/IAI.67.11.5768-5774.1999

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611. doi: 10.1038/nature13907

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

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222. doi: 10.1093/bioinformatics/btv710

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., et al. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* 12, 966–968. doi: 10.1038/nmeth.3505

Coscolla, M., and Gagneux, S. (2014). Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin. Immunol.* 26, 431–444. doi: 10.1016/j.smim.2014.09.012

Cousins, D. V., Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., et al. (2003). Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: *Mycobacterium pinnipedii* sp. nov. *Int. J. Syst. Evol. Microbiol.* 53(Pt 5), 1305–1314. doi: 10.1099/ijs.0.02401-0

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

Danilchanka, O., Sun, J., Pavlenok, M., Maueröder, C., Speer, A., Siroy, A., et al. (2014). An outer membrane channel protein of *Mycobacterium tuberculosis* with exotoxin activity. *Proc. Natl. Acad. Sci.* 111, 6750–6755. doi: 10.1073/pnas.1400136111

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., et al. (2017). Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *mBio* 8:e02133-16. doi: 10.1128/mBio.02133-16

Devis, M. C., Brites, D., Menardo, F., Loiseau, C. M., and Gagneux, S. (2020). Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microb Genom.* 7:000477. doi: 10.1099/mgen.0.000477

Fan, X., Abd Alla, A. A. E., and Xie, J. (2016). Distribution and function of prophage phiRv1 and phiRv2 among *Mycobacterium tuberculosis* complex. *J. Biomol. Struct. Dyn.* 34, 233–238. doi: 10.1080/07391102.2015.1022602

Garnier, T., Eiglmeier, K., Camus, J. C., Medina, N., Mansoor, H., Pryor, M., et al. (2003). The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl. Acad. Sci.* 100, 7877–7882. doi: 10.1073/pnas.1130426100

Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., and Cole, S. T. (1999). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.* 32, 643–655. doi: 10.1046/j.1365-2958.1999.01383.x

Heller, D., and Vingron, M. (2020). SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* 36, 5519–5521. doi: 10.1093/bioinformatics/btaa1034

Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W., Small, P. M. (2004). Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc. Natl. Acad. Sci.* 101, 4871–4876. doi: 10.1073/pnas.0305627101

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8:4061. doi: 10.1038/ncomms14061

Kiatichai, F., Xia, E., Hao, T. J., Yik-Ying, T., and Twee-Hee, O. R. (2016). In silico region of difference (RD) analysis of *Mycobacterium tuberculosis* complex from sequence reads using RD-Analyzer. *BMC Genomics* 17:847. doi: 10.1186/s12864-016-3213-1

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Meehan, C. J., Goig, G. A., Kohl, T. A., Verboven, L., and Rie, A. V. (2019). Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat. Rev. Microbiol.* 17, 533–545. doi: 10.1038/s41579-019-0214-5

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

Ngabonziza, J. C. S., Loiseau, C., Marceau, M., Jouet, A., Menardo, F., Tzfadia, O., et al. (2020). A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nat. Commun.* 11:2917. doi: 10.1038/s41467-020-16626-6

Niemann, S., Richter, E., and Rüschgerdes, S. (2002). Biochemical and genetic evidence for the transfer of *Mycobacterium tuberculosis* subsp. caprae Aranaz et al. 1999 to the species *Mycobacterium bovis* Karlson and Lessel 1970 (approved lists 1980) as *Mycobacterium bovis* subsp. caprae comb. nov. *Int. J. Syst. Evol. Microbiol.* 52, 433–436. doi: 10.1099/00207713-52-2-433

Palittapongarnpim, P., Ajawatanawong, P., Viratyosin, W., Smittipat, N., Disratthakit, A., Mahasirimongkol, S., et al. (2018). Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai *Mycobacterium tuberculosis* lineage 1 isolates. *Sci. Rep.* 8:11597. doi: 10.1038/s41598-018-29986-3

Parsons, S., Drewe, J. A., Pittius, N., Warren, R. M., and Helden, P. (2013). Novel cause of tuberculosis in meerkats South Africa. *Emerg. Infect. Dis.* 19, 2004–2007. doi: 10.3201/eid1912.130268

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* 12, 780–786. doi: 10.1038/nmeth.3454

Pittius, N., Helden, P., and Warren, R. M. (2012). Characterization of *Mycobacterium orygis*. *Emerg. Infect. Dis.* 18:1708. doi: 10.3201/eid1903.121005

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333–i339. doi: 10.1093/bioinformatics/bts378

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Sanou, A., Banuls, A. L., Godreuil, S., and Anh, N. (2015). *Mycobacterium tuberculosis*: ecology and evolution of a human bacterium. *J. Med. Microbiol.* 64, 1261–1269. doi: 10.1099/jmm.0.000171

Shitikov, E., Kolchenko, S., Mokrousov, I., Bespyatykh, J., Ischenko, D., Ilina, E., et al. (2017). Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. *Sci. Rep.* 7:10018. doi: 10.1038/s41598-017-10018-5

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374:abg8871. doi: 10.1126/science.abg8871

Soolingen, D. V., Hoogenboezem, T., Haas, P. D., Hermans, P., Koedam, M. A., Teppema, K. S., et al. (1997). A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex. Canetti: Characterization of an exceptional isolate from Africa. *Int. J. Syst. Bacteriol.* 47, 1236–1245. doi: 10.1099/00207713-47-4-1236

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

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

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

Tsolaki, A. G., Hirsh, A. E., DeRiemer, K., Enciso, J. A., Wong, M. Z., Hannan, M., et al. (2004). Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc. Natl. Acad. Sci.* 101, 4865–4870. doi: 10.1073/pnas.0305634101

Vasconcellos, S., Huard, R. C., Niemann, S., Kremer, K., and Ho, J. L. (2010). Distinct genotypic profiles of the two major clades of *Mycobacterium africanum*. *BMC Infect. Dis.* 10:80. doi: 10.1186/1471-2334-10-80

Vasimuddin, M., Misra, S., Li, H., et al. (2019). "Efficient architecture-aware acceleration of BWA-MEM for multicore systems," in *Proceedings of the 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, (Piscataway, NJ: IEEE), 314–324.

Wala, J. A., Bandopadhayay, P., Greenwald, N. F., O'Rourke, R., Sharpe, T., Stewart, C., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* 28, 581–591. doi: 10.1101/gr.221028.117

frontiers | Frontiers in Microbiology

TYPE Perspective
PUBLISHED 26 September 2022
DOI 10.3389/fmicb.2022.1021734

# Complete genome sequences of classical swine fever virus: Phylogenetic and evolutionary analyses

Yue Liu[1,2,3†], Amina Nawal Bahoussi[1†], Pei-Hua Wang[1], Changxin Wu[1,2,3] and Li Xing[1,2,3*]

[1]Institute of Biomedical Sciences, Shanxi University, Taiyuan, China, [2]The Key Laboratory of Medical Molecular Cell Biology of Shanxi Province, Shanxi University, Taiyuan, China, [3]Shanxi Provincial Key Laboratory for Prevention and Treatment of Major Infectious Diseases, Taiyuan, China

OPEN ACCESS

EDITED BY
Jianying Gu,
College of Staten Island, United States

REVIEWED BY
Xinglong Wang,
Northwest University, China

*CORRESPONDENCE
Li Xing
xingli107@gmail.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

RECEIVED 17 August 2022
ACCEPTED 02 September 2022
PUBLISHED 26 September 2022

CITATION
Liu Y, Bahoussi AN, Wang P-H, Wu C and Xing L (2022) Complete genome sequences of classical swine fever virus: Phylogenetic and evolutionary analyses. *Front. Microbiol.* 13:1021734. doi: 10.3389/fmicb.2022.1021734

The classical swine fever virus (CSFV) outbreaks cause colossal losses of pigs and drastic economic impacts. The current phylogenetic CSFV groups were determined mainly based on the partial genome. Herein, 203 complete genomic sequences of CSFVs collected worldwide between 1998 and 2018 available on the GenBank database were retrieved for re-genotyping and recombination analysis. The maximum likelihood phylogenetic tree determined two main groups, GI and GII, with multiple sub-genotypes. The "strain 39" (GenBank ID: AF407339), previously identified as belonging to sub-genotypes 1.1 or 2.2 based on the partial sequences, is found to be genetically distinct and independent, forming a new lineage depicted as GI-2.2b. Ten potential natural recombination events were identified, seven of which were collected in China and found involved in the genetic diversity of CSFVs. Importantly, the vaccine strains and highly virulent strains were all involved in the recombination events, which would induce extra challenges to vaccine development. These findings alarm that attenuated vaccines should be applied with discretion and recommend using subunit vaccines in parallel with other preventive strategies for better management of CSFVs.

KEYWORDS

**phylogenetic analysis, recombination, NS5 protein, attenuated vaccines, classical swine fever virus (CSFV)**

## Introduction

Classical swine fever (CSF), caused by CSF virus (CSFV), is a highly contagious disease affecting the *Suidae* family and causing massive pig production losses with severe global economic recession (Moennig, 2000). The clinical manifestations of CSF can range from acute to subacute and chronic forms depending on strain virulence and inducing various symptoms, including hyperthermia, anorexia, depression, vomiting, diarrhea, and skin hemorrhages (Moennig et al., 2003). CSF was first observed in the 1830s in Ohio in the United States of America, and since then, it has been reported globally, mainly in countries of America, Asia, and Eastern Europe (Edwards et al., 2000). In 1978, CSF was successfully eradicated in the USA; in 1963, Canada was free of CSF; the last case recorded in Chile was in 1996; Australia and New Zealand also achieved a CSF-free status in 1962 and 1953, respectively (Edwards et al., 2000). Although eradicated in many countries, CSFV is still endemic in the wild boar population, which causes sporadic outbreaks with an increased risk of re-emergence in domestic pigs, where direct and indirect transmissions have been demonstrated (Moennig, 2015; Xiang et al., 2017; Postel et al., 2019).

CSFV is a small enveloped virus containing a single-stranded, positive-sense RNA genome of approximately 12.3 kb. CSFV belongs to the *Pestivirus* genus in the *Flaviviridae* family (Thiel et al., 1991). The viral genome possesses a single open reading frame (ORF) of approximately 11.7 kb, which is flanked by a $5^{\prime}$-untranslated region ($5^{\prime}$-UTR), and a $3^{\prime}$-UTR and encodes a single polyprotein of ∼3,898 amino acids (Rümenapf et al., 1991, 1993; Thiel et al., 1991), that is processed during the virus replication into at least 12 functional units, including eight non-structural and four structural proteins. The non-structural protein (NSP) N-terminal protease ($N^{pro}$) is located at the N-terminus of the polyprotein, followed by the four structural proteins, *that is,* the core protein (C), envelope glycoprotein with RNase activity (Erns), and envelope glycoproteins E1 and E2. The remaining seven non-structural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B) are located at the C-terminus of the polyprotein (Thiel et al., 1991; Elbers et al., 1996). E2 glycoprotein and Erns are essential for virus attachment (Blome et al., 2017b). E2 could also form heterodimers with E1 and mediate the receptor binding and the virus entry through endocytosis (Shi et al., 2016).

CSF is a problematic issue for most countries of Asia. In China, CSF was first recognized in the 1920's (Tu et al., 2001), and the highly virulent strain (Shimen) was isolated in 1945 and used as a standard reference for vaccine evaluation. Since then, hundreds of outbreaks have been reported despite the immense efforts to contain the virus (Zhou, 2019). Eradication and elimination of CSFV infection in China, one of the major pig-producing countries, remains a constant challenge (Beer et al., 2015; Fan et al., 2021). Numerous studies revealed changes in the CSFV strains virulence (Shen et al., 2011; Ji et al., 2014; Zhang et al., 2018) and disease pathogenicity from acute to a chronic form, which was suggested to be related to vaccination (Ji et al., 2014).

Vaccination is the major control measure of CSFV in domestic pigs, and live-attenuated vaccines are widely applied in CSFV mandatory control programs in many countries. Those vaccines, including the Russian vaccine strain LK-VNIVViM, the Lapinized Philippines Coronel (LPC) strain, Chinese hog cholera lapinized virus (C-Strain), Japanese guinea-pig exaltation-negative (GPE-) strain, the Mexican PAV strains, and the French cell culture-adapted Thiverval strain, are developed by adaptive mutations after serial passages of main immunogen strains in rabbits or cell culture repeatedly (Baker, 1946; Koprowski et al., 1946; Oláh and Palatka, 1967; Lin et al., 1974; Ji et al., 2014; Coronado et al., 2021). Although live-attenuated vaccines are cost-effective and easy to use, particularly in pregnant sows and young pigs (van Oirschot, 2003), a strong controversy exists, such as in Europe, where domestic pigs are not vaccinated until severe outbreaks occur (Blome et al., 2017a).

CSFV isolates were initially classified into two genotypes based mainly on the complete E2 coding sequences (1119 nt) (Lowings et al., 1996; Postel et al., 2012), which is suggested as the most proper loci for more reliable phylogenetic analysis (Gong et al., 2016; Postel et al., 2019; Izzati et al., 2021; Zhu et al., 2021a). The complete E2 coding sequences are also recommended by the European Union (EU) and Office International des Epizooties (OIE) Reference Laboratory for reliable CSFV phylogenetic analysis (Postel et al., 2012). The global CSFV strains could also be classified into three genotypes and over thirteen subtypes (1.1–1.4, 2.1–2.5, and 3.1–3.4) mainly based on the E2 coding sequence (Zhou, 2019; Izzati et al., 2021; Singh et al., 2022). The correlation between field CSFV virulence and the evolutionary genotypes is not fully understood and not clearly established, although strains belonging to genotype 1 are the most highly virulent, while moderate or low virulent strains belong to genotypes 2 and 3 (Zhu et al., 2021b). Genotype 2 is the most prevalent genotype globally in recent years. Since the 1990's, the CSFV isolates in European countries belong to genotype 2 (2.1, 2.2, or 2.3) (Greiser-Wilke et al., 2000; Biagetti et al., 2001; Blome et al., 2010; Leifer et al., 2010; Postel et al., 2012; Simon et al., 2013) and are genetically distinct from genotype 1. All the American continent CSFV isolates belong to genotype 1. The Argentinian, Brazilian, Columbian, and Mexican isolates form four clusters in subgroup 1.1; the Honduran and Guatemalan CSFV strains are clustered in subgroup 1.3 (Zhou, 2019), and the Cuban isolates form a subgroup 1.4 (Postel et al., 2013). The CSFVs isolated in South Africa CSF outbreak in 2005 and in Israel in 2009 belong to subgroup 2.1 (Zhou, 2019). Several sub-genotypes 1.1, 2.1, 2.2, and 2.4 of CSFV isolates are reported in India, with 1.1 being dominant (Singh et al., 2022).

Since current phylogenetic classifications of CSFV are based mainly on partial genome sequences and may not provide sufficient information that helps fully understand the evolutionary character and genetic relatedness of the circulating strains, we reviewed in this report the full-length genomes of CSFVs by phylogenetical and recombination analysis.

## Re-genotyping of CSFV based on the complete genome sequences

A total of 203 complete genomic sequences of CSFVs, collected worldwide between 1977 and 2018 from 20 countries in Asia, Europe, and America, including China (66), South Korea (52), Japan (2), Mongolia (1), Vietnam (4), India (15), Germany (35), Lithuania (1), Denmark (6), Netherlands (2), Croatia (1), Bulgaria (1), Serbia (2), Spain (1), Russia (1), Sweden (1), Switzerland (7), France (1), USA (3), and Cuba (1), were retrieved from GenBank database. The maximum likelihood phylogenetic tree was constructed using MEGA-X software (Tamura and Nei, 1993; Tamura et al., 2021). The viruses in the current study were identified by their GenBank ID, name, country, and year of collection.

As shown in Figure 1, Supplementary Table 1, and Supplementary Figures 1–3, the CSFV full-length genome sequences cluster into two main groups, GI and GII. The GI group includes 95 CSFV strains and contains five sub-genotypes: 1.1, 1.2, 2.2b, 3.2, and 3.4; meanwhile GII group includes 108 CSFV strains and contains five sub-genotypes: 2.1a, 2.1b, 2.1c, 2.2, and 2.3. We identified a mixed population of CSFV genotypes and sub-genotypes co-circulating in China, Germany, and South Korea, where the earliest strains fall into GI while the more recently collected strains fall into GII except South Korea CSFV strains. Interestingly, the GI-2.2b is a new sub-genotype identified based on the full-length genome phylogenetic analysis and is found restricted to China. These findings corroborate the previous genotyping based on the partial genomic sequences, such as 5′-UTR and E2 gene et al., where most of the strains included in this study fall into their corresponding sub-genotypes. For example, the strain CSF1048 (GenBank ID: HQ148063) assigned in 2.1b based on the complete E2 genomic sequences (Leifer et al., 2011) clusters into 2.1b based on the full-length genome and is genetically close to China strains GXWZ02 (GenBank ID: AY367767) (Li et al., 2006), SXYL2006 (GenBank ID: GQ122383) (Li et al., 2011), Zj0801 (GenBank ID: FJ529205), HEBZ (GenBank ID: GU592790), and 0406/CH (GenBank ID: AY568569). Similarly, Paderborn (GenBank ID: AY072924), assigned to 2.1a based on the partial E2 sequences (Uttenthal et al., 2001), clusters in 2.1a in the GII based on the full-length genome and is genetically close to two China strains SXCDK (GenBank ID: GQ923951) and 96TD (GenBank ID: AY554397). Viruses HNLY-2011 (GenBank ID: JX262391) and HNSD-2012 (GenBank ID: JX218094) were reported as a new sub-genotype 2.1c (Jiang et al., 2013) and genetically close to the virus HY78 (GenBank ID: MH979231.1) isolated in Vietnam in 2015 (Kim et al., 2019). The first identified sub-genotype 2.2 virus LAL-290 (GenBank ID: KC851953) (Kumar et al., 2014) still forms an independent lineage 2.2 in group GII with other genetically closer genomes. The reference strain Alfort/Tuebingen (GenBank ID: J04358) assigned to 2.3 based on 5′-UTR, or partial E2 gene (Meyers et al., 1989), and Uelzen (GenBank ID: GU324242), Euskirchen (GenBank ID: GU233732), Borken (GenBank ID: GU233731), Roesrath (GenBank ID: GU233734), and Hennef (GenBank ID: GU233733) assigned to 2.3 based on complete genome, 5′-UTR, N$^{pro}$, or E2 gene (Leifer et al., 2010) cluster into the same sub-genotype 2.3 in this study. The highly virulent strain Brescia (GenBank ID: M31768), assigned to 1.2 based on 5′-UTR, or partial E2 gene (Moormann et al., 1990), fall into 1.2 based on the full-length genome sequences and is genetically close to the attenuated vaccine strain RUCSFPLUM (GenBank ID: AY578688) isolated in 2001 (Risatti et al., 2005).

CSFV strains, P97 (GenBank ID: L49347) (Shiu et al., 1996) and TW-94 (GenBank ID: AY646427), were reported as a new sub-genotype 3.4 based on the complete genome sequences (Lin et al., 2007); YI9908 (GenBank ID: KT716271) and JJ9811 (GenBank ID: KF669877.1) isolated in South Korea form an independent sub-genotype 3.2 (Lim et al., 2016). Both 3.4 and 3.2 fall into the GI group, genetically closer to 1.2 and 1.1. The sub-genotype 1.1 could be further divided into 1.1a, 1.1b, and 1.1c based on the full-genome sequences in this analysis. The highly virulent strain Koslov (GenBank ID: HM237795), previously assigned as 1.1 based on complete genome, 5′-NTR, N$^{pro}$, or E2 protein sequences (Leifer et al., 2010), clusters into 1.1a with the strain HCLV (GenBank ID: AF091507). The NG79-11 (GenBank ID: KC503764) (Tomar et al., 2015) and VB-131 (GenBank ID: KM262189) (Kamboj et al., 2014) isolated in India are assigned as 1.1b together with Shimen strain (GenBank ID: AF092448), a highly virulent strain isolated in China in 1945 (Zhang et al., 2018), SWH (GenBank ID: DQ127910) (Li et al., 2006), and GZ-2009 (GenBank ID: HQ380231) (Shen et al., 2011). Alfort/187 (GenBank ID: X87939) (Ruggli et al., 1996) was under 1.1c together with the high virulent strains Glentorf (GenBank ID: U45478), CAP (GenBank ID: X96550), and Alfort A19 (GenBank ID: U90951), and the most of viruses isolated in South Korea during 1987–2019.

Compared to partial genomic sequence-based analysis, the complete genome sequence-based phylogenetic analysis would provide a more accurate understanding of the genetic relatedness of CSFV “Strain 39” (GenBank ID: AF407339) belonging to sub-genotypes 1.1 or 2.2 based on the 5’-UTR fragment or the entire 5’-UTR-E2 sequences, respectively (Postel et al., 2012), was recently indicated to be genetically closer to the Indian strain CSFV IND/UK/LAL-290 (GenBank ID: KC851953) in 2.2 based on complete E2 sequences (Zhu et al., 2021b). Herein, the full-length genomic sequence revealed that “Strain 39” is genetically distinct from viruses of 1.1 or 2.2 sub-genotypes and forms a new independent lineage depicted as GI-2.2b in this analysis.

A group of viruses including JSZL (GenBank ID: KT119352), SDSG1410 (GenBank ID: MF150645), SDLS1410 (GenBank ID: MF150644) (Zhang et al., 2015), HB150309 (GenBank ID: MF150640), JL150418 (GenBank ID: MF150642), NK150425 (GenBank ID: MF150643), SDZC150601 (GenBank ID: MF150646), HLJ1 (GenBank ID: MF150641), and SDWF-2016 (GenBank ID: MK211486) (Zhang et al., 2018) were claimed as a new sub-genotype 2.1d based on full or partial E2 sequences. However, according to Figure 1, these 2.1d sub-genotype isolates are genetically within the 2.1b sub-genotype and do not form a distinct lineage as reported before. As shown in Figure 1 and

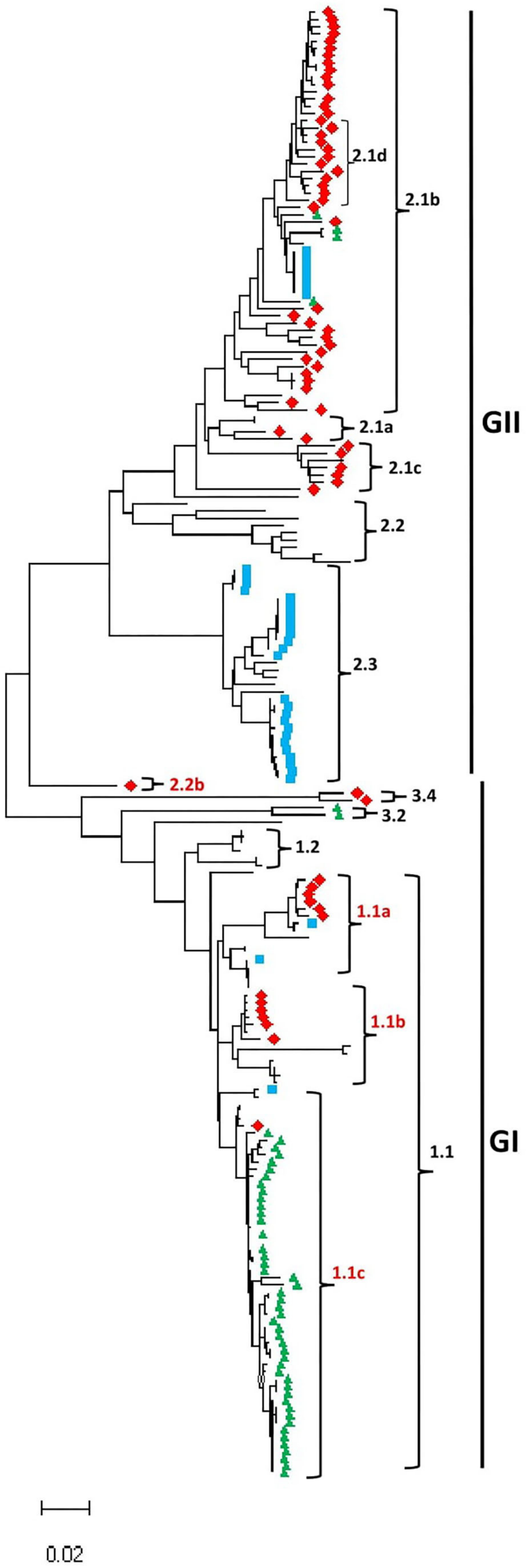


FIGURE 1
Maximum likelihood phylogenetic tree based on the full-length genome of CSFV isolated worldwide from 1998 to 2018
*(Continued)*

FIGURE 1 (Continued)
(continue in Supplementary Figures 1–3). Multiple-sequence alignments were performed using Clustal Omega server, and the phylogenetic tree was constructed from the aligned nucleotide sequences using the maximum likelihood method in the MEGA-X software. The numbers at each branch represent bootstrap values of 1,000 replicates. The scale bars indicate the number of inferred substitutions per site. Viruses isolated in China are indicated with red diamonds, South Korea in green triangles, and Germany in blue squares.

Supplementary Table 1, the China strains could be found in most of sub-genotypes.

To increase the stringency and reliability of our phylogenetic tree findings, we performed a similarity analysis, comparing the complete genome of HEBZ (GenBank ID: GU592790) in 2.1b sub-genotype to 13 representative CSFV full-length sequences for each sub-genotypes using SimPlot analysis (Lole et al., 1999). As shown in Supplementary Figure 4, the E1, E2, NS2, NS4, and NS5 coding regions revealed low similarity levels, where strains from GI exhibited the lowest percentage levels (< 80%) and are shown distant from those of GII that exhibited higher similarity percentage (>90%). However, the same genomic fragments displayed distinct similarity levels among strains of GII group ranging between 90 and 98%, contrary to strains of GI group (Supplementary Figure 4). The latter indicates the divergences between CSFV sub-genotypes and corroborates the high diversity of GII group (Figure 1). Interestingly, "strain 39" (GenBank ID: AF407339), which formed a new lineage 2.2b in GI, resembled highly GII before NS4B coding region, but become closer to GI in NS5 coding region. These findings are consistent with the phylogenetic analysis results, suggesting that the defined CSFV genotypes and sub-genotypes are distinctly shown and highly specific, with a significant difference in the genome between GI and GII (Supplementary Figure 4).

## Genomic recombination between the worldwide collected CSFVs

To understand the mechanisms behind the adaptation and genomic diversity of the worldwide circulating CSFVs, we conducted a recombination analysis of 203 complete genomic sequences of CSFV using the seven algorithms of the RDP4 software package (Martin et al., 2015). Our recombination analysis identified 10 natural recombination events (Supplementary Table 2), six of which occurred between GII strains, three occurred between GI strains, and only one was intergenotype (GI vs GII) (Supplementary Table 2, Figure 2). Importantly, most of the identified recombinant strains were collected in China (Events 1, 4, 5, 6, 7, 8, and 9), four events among which occurred between China strains (Events 4, 6, 7, and 8), two between China and Germany strains (Events 5 and

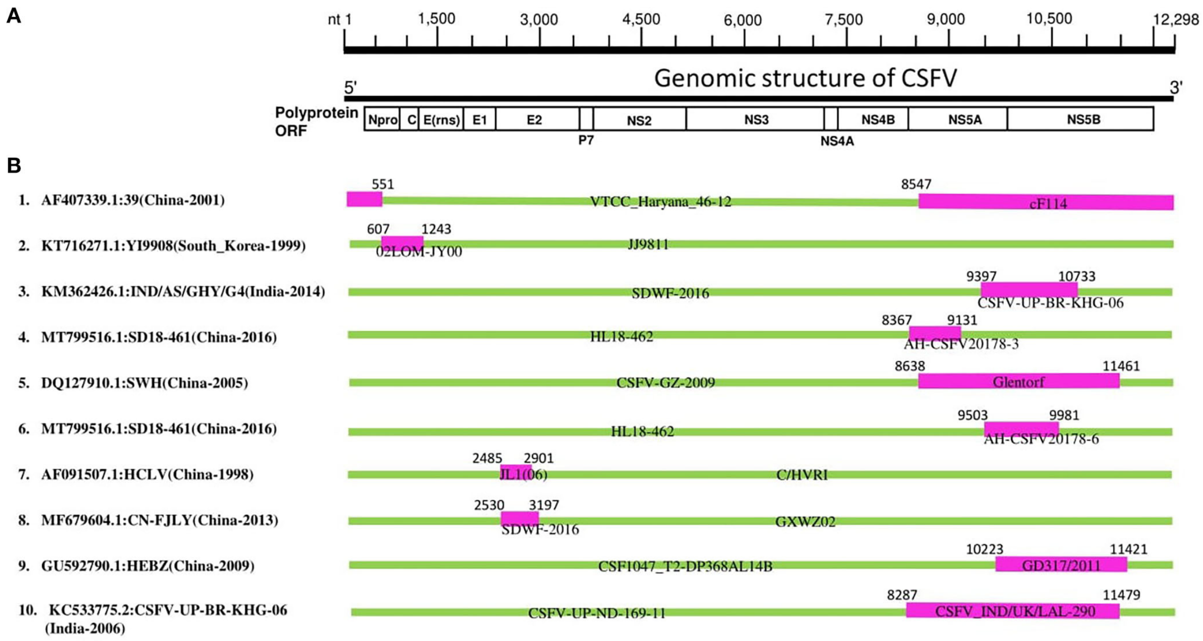


FIGURE 2

CSFV complete genome recombination. **(A)** Diagram showing the full-length genome of CSFV strain GD 19/2011 (GenBank ID: KU 504339) and the corresponding regions encoding $N^{pro}$ C, Erns, E1, E2, p7 N2, NS3, NS4A, NS4B, NS5A, and NS5B. **(B)** Schematic representation of potential recombination events listed in Supplementary Table 2. The recombination event serial number and the description of potential recombinants (GenBank ID: virus name/country-collection year) are shown on the left. The filled pink and green blocks represent the genomic regions from minor or major parent viruses, respectively. The numbers on the top of filled green blocks indicate the nucleotide positions of breakpoints on the corresponding recombinant viruses.

9), and one occurred between China and India strains (Event 1) (Supplementary Table 2, Figure 2). Furthermore, recombination occurred between South Korean strains (Event 2) and between Indian strains (Event 10) (Supplementary Table 2, Figure 2).

In previous studies (He et al., 2007; Chen and Chen, 2014), the "strain 39" (GenBank ID: AF407339) isolated in China has been identified to stem from nature recombination. Consistently, "strain 39" in our report is identified as a recombinant strain (Event 1) resulting from natural recombination between cF114 (GenBank ID: AF333000.1) and VTCC_Haryana_46-12 (GenBank ID: MK405702) as minor and major parental sequences, respectively, which is supported by similarity analysis where the genome of "strain 39" highly resembled GII before NS4B region, but become closer to GI in NS5 region (Supplementary Figure 4), and also the recombination map where the beginning breakpoint was located at nt 8547 corresponding to NS5 fragment (Figure 2).

Furthermore, Lim et al. reported that the YI9908 strain shares with the JJ9811 strain a 95.7% homology at the nucleotide (nt) level and 95.6% at the amino acid level (Lim et al., 2016). Comparative analysis of YI9908 and JJ9811 strains revealed a low nucleotide sequence homology for the $N^{pro}$ gene (90.1%) and the C gene (87.5%) (Lim et al., 2016). Consistently, our analysis identified YI9908 (GenBank ID: KT716271) as a recombinant strain (Event 2) and the JJ9811 strain (GenBank ID: KF669877) as its major parent (Supplementary Table 2), and the recombination map exhibited the beginning and ending breakpoints at nt 607-1243, corresponding to the $N^{pro}$ and the C genes (Supplementary Figure 5).

It has been reported recently that Chinese CSFV HCLV (AF091507.1) resulted from natural recombination between Shimen (AF092448.2) and CSFV strain C/HVRI (AY805221.1) with two recombination breakpoints at nt 2484 and 2900. Similarly, in our report, HCLV (AF091507.1) is identified as a recombinant strain (Event 7) with identical breakpoints location at nt 2485-2901; however, HCLV is determined resulting from a different parental strain: JL1(06) (EU497410.1) with C/HVRI (AY805221.1). These findings suggest genetic mosaicism of the HCLV virus (Supplementary Table 2, Figure 2).

Importantly, the highly virulent and the vaccine strains are all identified as involved in recombination during the genetic evolution of CSFVs. Our phylogenetic tree revealed that recombinant YI9908 (Event 2) and its major parent

JJ9811 are genetically close to the vaccine strains: LK-VNIVViM and Rovac, while its minor parent 02LOM-JY00 is itself a vaccine strain applied in South Korea (Supplementary Figure 3). The recombinant HCLV (Event 7) and its major parent C/HVRI are all vaccine strains used in China, whereas the minor parent JL1(06) (GenBank ID: EU497410) is a highly virulent strain (Supplementary Figure 3). Therefore, the high rate of recombination, identified between China CSFV strains, particularly in the GII-2 group, is a significant threat to the pork industry and might be related to the applied vaccination programs, which should be merited particular attention.

As shown in Figure 2, in eight out of ten CSFV recombination events (Asia strains), the beginning and ending breakpoints are found to be within 5'- or 3'- proximal regions of the genome, encoding the $N^{pro}$, C, NS5A, and NS5B proteins, respectively. The NS5A protein has been demonstrated to regulate the CSFV replication and viral RNA synthesis by interacting with NS5B and $3^{'}$-UTR, where low levels of NS5A stimulate the virus replication while high levels of NS5A suppress the RNA replication (Xu et al., 2020). Therefore, these recombination characteristics partly explain the adaptation and evolution of the circulating CSFV strains and predict future severe outbreaks that might challenge vaccine development, where subunit vaccines are highly recommended instead of attenuated forms.

We further verified the authenticity of the identified recombination events by building phylogenetic trees based on different genomic fragments of CSFV (relative to the strain GD19/2011, GenBank: KU504339.1) (Supplementary Figure 5): The fragment nt 6000-7000 encodes NS3 protein, while nt 9500-10000 encodes part of NS5A and NS5B proteins. As shown in Supplementary Figure 5, the phylogenetic trees based on two genome fragments were not superimposed. Notably, the recombinant strains (indicated in red color) of each event are genetically closer to their minor parents (indicated in blue color) in the phylogenetic tree based on nt 9,500–10,000 (Supplementary Figure 5) and become closer to their major parents (indicated with yellow color) in the phylogenetic tree based on nt 6,000-7,000 (Supplementary Figure 5). These results are congruent with the recombination mapping, indicating that the identified recombination events are real.

Therefore, this report exhibited that natural recombination was driving the genetic diversity and complexity of circulating CSFV strains. As the attenuated vaccines were found involved in the recombination events, their application should be cautious, and subunit vaccines are highly recommended for more successful control and effective prevention.

## 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

AB, YL, and LX conceived the study and revised the manuscript. AB, YL, and P-HW performed analysis. YL and AB wrote the manuscript. LX and CW supervised analysis. All authors read and approved the final manuscript.

## Funding

This work is funded by the Programme of Introducing Talents of Discipline to Universities (D21004) and the Fundamental Research Program of Shanxi Province, China (20210302124187).

## Acknowledgments

We thank all the persons who contributed to the collection and production of CSFV genomic sequences in GenBank.

## 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.2022.1021734/full#supplementary-material

## References

Baker, J. A. (1946). Serial passage of hog cholera virus in rabbits. *Proc. Soc. Exp. Biol. Med.* 63, 183–187. doi: 10.3181/00379727-63-15541

Beer, M., Goller, K. V., Staubach, C., and Blome, S. (2015). Genetic variability and distribution of classical swine fever virus. *Anim. Health Res. Rev.* 16, 33–39. doi: 10.1017/S1466252315000109

Biagetti, M., Greiser-Wilke, I., and Rutili, D. (2001). Molecular epidemiology of classical swine fever in Italy. *Vet. Microbiol.* 83, 205–215. doi: 10.1016/S0378-1135(01)00424-2

Blome, S., Grotha, I., Moennig, V., and Greiser-Wilke, I. (2010). Classical swine fever virus in South-Eastern Europe–retrospective analysis of the disease situation and molecular epidemiology. *Vet. Microbiol.* 146, 276–284. doi: 10.1016/j.vetmic.2010.05.035

Blome, S., Moß, C., Reimann, I., König, P., and Beer, M. (2017a). Classical swine fever vaccines-State-of-the-art. *Vet. Microbiol.* 206, 10–20. doi: 10.1016/j.vetmic.2017.01.001

Blome, S., Staubach, C., Henke, J., Carlson, J., and Beer, M. (2017b). Classical swine fever-an updated review. *Viruses* 9. 86. doi: 10.3390/v9040086

Chen, Y., and Chen, Y. F. (2014). Extensive homologous recombination in classical swine fever virus: a re-evaluation of homologous recombination events in the strain AF407339. *Saudi J. Biol. Sci.* 21, 311–316. doi: 10.1016/j.sjbs.2013.12.004

Coronado, L., Perera, C. L., Rios, L., Frías, M. T., and Pérez, L. J. (2021). A critical review about different vaccines against classical swine fever virus and their repercussions in endemic regions. *Vaccines* 9, 154. doi: 10.3390/vaccines9020154

Edwards, S., Fukusho, A., Lefèvre, P.-C., Lipowski, A., Pejsak, Z., Roehe, P., et al. (2000). Classical swine fever: the global situation. *Vet. Microbiol.* 73, 103–119. doi: 10.1016/S0378-1135(00)00138-3

Elbers, K., Tautz, N., Becher, P., Stoll, D., Rümenapf, T., and Thiel, H. J. (1996). Processing in the pestivirus E2-NS2 region: identification of proteins p7 and E2p7. *J. Virol.* 70, 4131–4135. doi: 10.1128/jvi.70.6.4131-4135.1996

Fan, J., Liao, Y., Zhang, M., Liu, C., Li, Z., Li, Y., et al. (2021). Anti-classical swine fever virus strategies. *Microorganisms* 9, 761. doi: 10.3390/microorganisms9040761

Gong, W., Wu, J., Lu, Z., Zhang, L., Qin, S., Chen, F., et al. (2016). Genetic diversity of subgenotype 2.1 isolates of classical swine fever virus. *Infect. Genet. Evol.* 41, 218–226. doi: 10.1016/j.meegid.2016.04.002

Greiser-Wilke, I., Fritzemeier, J., Koenen, F., Vanderhallen, H., Rutili, D., De Mia, G. M., et al. (2000). Molecular epidemiology of a large classical swine fever epidemic in the European Union in 1997-1998. *Vet. Microbiol.* 77, 17–27. doi: 10.1016/S0378-1135(00)00253-4

He, C. Q., Ding, N. Z., Chen, J. G., and Li, Y. L. (2007). Evidence of natural recombination in classical swine fever virus. *Virus Res.* 126, 179–185. doi: 10.1016/j.virusres.2007.02.019

Izzati, U. Z., Hoa, N. T., Lan, N. T., Diep, N. V., Fuke, N., Hirai, T., et al. (2021). Pathology of the outbreak of subgenotype 2.5 classical swine fever virus in northern *Vietnam.* 7, 164–174. doi: 10.1002/vms3.339

Ji, W., Niu, D. D., Si, H. L., Ding, N. Z., and He, C. Q. (2014). Vaccination influences the evolution of classical swine fever virus. *Infect. Genet. Evol.* 25, 69–77. doi: 10.1016/j.meegid.2014.04.008

Jiang, D. L., Liu, G. H., Gong, W. J., Li, R. C., Hu, Y. F., Tu, C., et al. (2013). Complete genome sequences of classical Swine Fever virus isolates belonging to a new subgenotype, 2.1c, from Hunan province, China. *Genome Announc.* 1, e00080-12. doi: 10.1128/genomeA.00080-12

Kamboj, A., Patel, C. L., Chaturvedi, V. K., Saini, M., and Gupta, P. K. (2014). Complete genome sequence of an Indian field isolate of classical swine fever virus belonging to subgenotype 1.1. *Genome Announc.* 2, e00886-14. doi: 10.1128/genomeA.00886-14

Kim, K. S., Le, V. P., Choe, S., Cha, R. M., Shin, J., Cho, I. S., et al. (2019). Complete genome sequences of classical swine fever virus subgenotype 2.1 and 2.2 strains isolated from vietnamese pigs. *Microbiol Resour Announc.* 8, e01634-18. doi: 10.1128/MRA.01634-18

Koprowski, H., James, T. R., and Cox, H. R. (1946). Propagation of hog cholera virus in rabbits. *Proc. Soc. Exp. Biol. Med.* 63, 178–183. doi: 10.3181/00379727-63-15540

Kumar, R., Rajak, K. K., Chandra, T., Thapliyal, A., Muthuchelvan, D., Sudhakar, S. B., et al. (2014). Whole-genome sequence of a classical swine fever virus isolated from the uttarakhand state of India. *Genome Announc.* 2, e00371-14. doi: 10.1128/genomeA.00371-14

Leifer, I., Hoeper, D., Blome, S., Beer, M., and Ruggli, N. (2011). Clustering of classical swine fever virus isolates by codon pair bias. *BMC Res. Notes* 4, 521. doi: 10.1186/1756-0500-4-521

Leifer, I., Hoffmann, B., Höper, D., Bruun Rasmussen, T., Blome, S., Strebelow, G., et al. (2010). Molecular epidemiology of current classical swine fever virus isolates of wild boar in Germany. *J. Gen. Virol.* 91, 2687–2697. doi: 10.1099/vir.0.023200-0

Li, L. W., Zhang, Z., Wu, F. X., Cheng, W., Yang, R. S., Zhang, Y. X., et al. (2011). Cloning and analysis of the complete genome of CSFV isolate SXCDK from Shaanxi province. *J. Northwest A & F Univ.* 39, 1–7. Available online at: http://www.xnxbz.net/xbnlkjdxzren/ch/reader/create_pdf.aspx?file_no=20110101&flag=1&year_id=2011&quarter_id=1

Li, X., Xu, Z., He, Y., Yao, Q., Zhang, K., Jin, M., et al. (2006). Genome comparison of a novel classical swine fever virus isolated in China in 2004 with other CSFV strains. *Virus Genes* 33, 133–142. doi: 10.1007/s11262-005-0048-2

Lim, S. I., Han, S. H., Hyun, H., Lim, J. A., Song, J. Y., Cho, I. S., et al. (2016). Complete genome sequence analysis of acute and mild strains of classical swine fever virus subgenotype 3.2. *Genome Announc.* 4, e01329-15. doi: 10.1128/genomeA.01329-15

Lin, T. C., Shieh, C. M., and Su, J. F. (1974). Virus multiplication in pigs inoculated with lapinized hog cholera live vaccine. *Chin. J. Microbiol.* 7, 13–19.

Lin, Y. J., Chien, M. S., Deng, M. C., and Huang, C. C. (2007). Complete sequence of a subgroup 3.4 strain of classical swine fever virus from Taiwan. *Virus Genes* 35, 737–744. doi: 10.1007/s11262-007-0154-4

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

Lowings, P., Ibata, G., Needham, J., and Paton, D. (1996). Classical swine fever virus diversity and evolution. *J Gen Virol.* 77, 1311–1321. doi: 10.1099/0022-1317-77-6-1311

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1, vev003. doi: 10.1093/ve/vev003

Meyers, G., Rümenapf, T., and Thiel, H. J. (1989). Molecular cloning and nucleotide sequence of the genome of hog cholera virus. *Virology* 171, 555–567. doi: 10.1016/0042-6822(89)90625-9

Moennig, V. (2000). Introduction to classical swine fever: virus, disease and control policy. *Vet. Microbiol.* 73, 93–102. doi: 10.1016/S0378-1135(00)00137-1

Moennig, V. (2015). The control of classical swine fever in wild boar. *Front. Microbiol.* 6, 1211. doi: 10.3389/fmicb.2015.01211

Moennig, V., Floegel-Niesmann, G., and Greiser-Wilke, I. (2003). Clinical signs and epidemiology of classical swine fever: a review of new knowledge. *Vet. J.* 165, 11–20. doi: 10.1016/S1090-0233(02)00112-0

Moormann, R. J., Warmerdam, P. A., van der Meer, B., Schaaper, W. M., Wensvoort, G., and Hulst, M. M. (1990). Molecular cloning and nucleotide sequence of hog cholera virus strain Brescia and mapping of the genomic region encoding envelope protein E1. *Virology* 177, 184–198. doi: 10.1016/0042-6822(90)90472-4

Oláh, P., and Palatka, Z. (1967). Immunobiological study of lapinized hog cholera virus strains. *Acta Vet. Acad. Sci. Hung.* 17, 239–247.

Postel, A., Nishi, T., Kameyama, K. I., Meyer, D., Suckstorff, O., Fukai, K., et al. (2019). Reemergence of classical swine fever, Japan, 2018. *Emerging Infect. Dis.* 25, 1228–1231. doi: 10.3201/eid2506.181578

Postel, A., Schmeiser, S., Bernau, J., Meindl-Boehmer, A., Pridotkas, G., Dirbakova, Z., et al. (2012). Improved strategy for phylogenetic analysis of classical swine fever virus based on full-length E2 encoding sequences. *Vet. Res.* 43, 50. doi: 10.1186/1297-9716-43-50

Postel, A., Schmeiser, S., Perera, C. L., Rodríguez, L. J., Frias-Lepoureau, M. T., and Becher, P. (2013). Classical swine fever virus isolates from Cuba form a new subgenotype 1.4. *Vet. Microbiol.* 161, 334–338. doi: 10.1016/j.vetmic.2012.07.045

Risatti, G. R., Borca, M. V., Kutish, G. F., Lu, Z., Holinka, L. G., French, R. A., et al. (2005). The E2 glycoprotein of classical swine fever virus is a virulence determinant in swine. *J. Virol.* 79, 3787–3796. doi: 10.1128/JVI.79.6.3787-3796.2005

Ruggli, N., Tratschin, J. D., Mittelholzer, C., and Hofmann, M. A. (1996). Nucleotide sequence of classical swine fever virus strain Alfort/187 and transcription of infectious RNA from stably cloned full-length cDNA. *J. Virol.* 70, 3478–3487. doi: 10.1128/jvi.70.6.3478-3487.1996

Rümenapf, T., Meyers, G., Stark, R., and Thiel, H. J. (1991). Molecular characterization of hog cholera virus. *Arch. Virol. Suppl.* 3, 7–18. doi: 10.1007/978-3-7091-9153-8_2

Rümenapf, T., Unger, G., Strauss, J. H., and Thiel, H. J. (1993). Processing of the envelope glycoproteins of pestiviruses. *J. Virol.* 67, 3288–3294. doi: 10.1128/jvi.67.6.3288-3294.1993

Shen, H., Pei, J., Bai, J., Zhao, M., Ju, C., Yi, L., et al. (2011). Genetic diversity and positive selection analysis of classical swine fever virus isolates in south China. *Virus Genes* 43, 234–242. doi: 10.1007/s11262-011-0625-5

Shi, B. J., Liu, C. C., Zhou, J., Wang, S. Q., Gao, Z. C., Zhang, X. M., et al. (2016). Entry of classical swine fever virus into PK-15 Cells via a pH-, dynamin-, and cholesterol-dependent, clathrin-mediated endocytic pathway that requires Rab5 and Rab7. *J. Virol.* 90, 9194–9208. doi: 10.1128/JVI.00688-16

Shiu, J. S., Chang, M. H., Liu, S. T., Ho, W. C., Lai, S. S., Chang, T. J., et al. (1996). Molecular cloning and nucleotide sequence determination of three envelope genes of classical swine fever virus Taiwan isolate p97. *Virus Res.* 41, 173–178. doi: 10.1016/0168-1702(96)01286-5

Simon, G., Le Dimna, M., Le Potier, M. F., and Pol, F. (2013). Molecular tracing of classical swine fever viruses isolated from wild boars and pigs in France from 2002 to 2011. *Vet. Microbiol.* 166, 631–638. doi: 10.1016/j.vetmic.2013.06.032

Singh, N., Batra, K., Chaudhary, D., Punia, M., Kumar, A., Maan, N. S., et al. (2022). Prevalence of porcine viral respiratory diseases in India. *Anim. Biotechnol.* 3, 1–13. doi: 10.1080/10495398.2022.2032117

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.

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

Thiel, H. J., Stark, R., Weiland, E., Rümenapf, T., and Meyers, G. (1991). Hog cholera virus: molecular composition of virions from a pestivirus. *J. Virol.* 65, 4705–4712. doi: 10.1128/jvi.65.9.4705-4712.1991

Tomar, N., Gupta, A., Arya, R. S., Somvanshi, R., Sharma, V., and Saikumar, G. (2015). Genome sequence of classical Swine Fever virus genotype 1.1 with a genetic marker of attenuation detected in a continuous porcine cell line. *Genome Announc.* 3, e00375-15 doi: 10.1128/genomeA.00375-15

Tu, C., Lu, Z., Li, H., Yu, X., Liu, X., Li, Y., et al. (2001). Phylogenetic comparison of classical swine fever virus in China. *Virus Res.* 81, 29–37. doi: 10.1016/S0168-1702(01)00366-5

Uttenthal, A., Le Potier, M. F., Romero, L., De Mia, G. M., and Floegel-Niesmann, G. (2001). Classical swine fever (CSF) marker vaccine. Trial I. Challenge studies in weaner pigs. *Vet. Microbiol.* 83, 85–106. doi: 10.1016/S0378-1135(01)00409-6

van Oirschot, J. T. (2003). Vaccinology of classical swine fever: from lab to field. *Vet. Microbiol.* 96, 367–384. doi: 10.1016/j.vetmic.2003.09.008

Xiang, H., Gao, J., Cai, D., Luo, Y., Yu, B., Liu, L., et al. (2017). Origin and dispersal of early domestic pigs in northern China. *Sci. Rep.* 7, 5602. doi: 10.1038/s41598-017-06056-8

Xu, C., Feng, L., Chen, P., Li, A., Guo, S., Jiao, X., et al. (2020). Viperin inhibits classical swine fever virus replication by interacting with viral nonstructural 5A protein. *J. Med. Virol.* 92, 149–160. doi: 10.1002/jmv.25595

Zhang, H., Leng, C., Feng, L., Zhai, H., Chen, J., Liu, C., et al. (2015). A new subgenotype 2.1d isolates of classical swine fever virus in China, 2014. *Infect. Genet. Evol.* 34, 94–105. doi: 10.1016/j.meegid.2015.05.031

Zhang, H., Leng, C., Tian, Z., Liu, C., Chen, J., Bai, Y., et al. (2018). Complete genomic characteristics and pathogenic analysis of the newly emerged classical swine fever virus in China. *BMC Vet. Res.* 14, 204. doi: 10.1186/s12917-018-1504-2

Zhou, B. (2019). Classical swine fever in China-an update minireview. *Front. Vet. Sci.* 6, 187. doi: 10.3389/fvets.2019.00187

Zhu, E., Wu, H., Chen, W., Qin, Y., Liu, J., Fan, S., et al. (2021a). Classical swine fever virus employs the PERK- and IRE1-dependent autophagy for viral replication in cultured cells. *Virulence* 12, 130–149. doi: 10.1080/21505594.2020.1845040

Zhu, X., Liu, M., Wu, X., Ma, W., and Zhao, X. (2021b). Phylogenetic analysis of classical swine fever virus isolates from China. *Arch. Virol.* 166, 2255–2261. doi: 10.1007/s00705-021-05084-0

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fmicb.2022.1012124

OPEN ACCESS

EDITED BY
Zhan Zhou,
Zhejiang University,
China

REVIEWED BY
Lin Wei,
Soochow University,
China
Feng He,
Capital Institute of Pediatrics, China

*CORRESPONDENCE
Hongxing Shen
hxshen@ujs.edu.cn
Hua Wang
wangjiahan1979@163.com

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

SPECIALTY SECTION
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

RECEIVED 05 August 2022
ACCEPTED 20 September 2022
PUBLISHED 13 October 2022

CITATION
Liu T, Li Y, Chen S, Wang L, Liu X, Yang Q, Wang Y, Qiao X, Tong J, Deng X, Shao S, Wang H and Shen H (2022) CircDDX17 enhances coxsackievirus B3 replication through regulating miR-1248/NOTCH receptor 2 axis.
*Front. Microbiol.* 13:1012124.
10.3389/fmicb.2022.1012124

# CircDDX17 enhances coxsackievirus B3 replication through regulating miR-1248/NOTCH receptor 2 axis

Tingjun Liu[1,2†], Yuhan Li[1,2†], Shengjie Chen[1†], Lulu Wang[2], Xiaolan Liu[2], Qingru Yang[2], Yan Wang[2], Xiaorong Qiao[2], Jing Tong[3], Xintao Deng[4], Shihe Shao[1], Hua Wang[2*] and Hongxing Shen[1,2*]

[1]Cardiothoracic Surgery, Affiliated Hospital of Jiangsu University, Zhenjiang, China, [2]Department of Laboratory Medicine, School of Medicine, Jiangsu University, Zhenjiang, China, [3]Xuzhou Center for Disease Control and Prevention, Xuzhou, China, [4]People's Hospital of Xinghua, Jiangsu University Teaching Hospital, Xinghua, China

Coxsackievirus B3 (CVB3) was one of the most common pathogens to cause viral myocarditis. Circular RNAs as novel non-coding RNAs with a closed loop molecular structure have been confirmed to be involved in virus infectious diseases, but the function in CVB3 infection was not systematically studied. In this study, we identified that hsa_circ_0063331 (circDDX17) was drastically decreased after CVB3 infection by circRNA microarray. *In vivo* and *in vitro*, when cells or mice were infected with CVB3, the expression of circDDX17 was significantly reduced, as demonstrated by quantitative real-time PCR assays. Additionally, circDDX17 enhanced CVB3 replication by downregulating the expression of miR-1248 in HeLa and HL-1 cells, and miR-1248 regulated CVB3 replication through interacting with the gene coding for NOTCH Receptor 2 (NOTCH2), and NOTCH2 could upregulate methyltransferase-like protein 3 (METTL3). Taken together, this study suggested that circDDX17 promoted CVB3 replication and regulated NOTCH2 by targeting miR-1248 as a miRNAs sponge.

KEYWORDS

**coxsackievirus B3, circular RNAs, miRNAs, NOTCH2, METTL3**

## Introduction

Coxsackievirus B3 (CVB3) is an RNA virus belonging to the *Enterovirus* genus of *Picornaviridae*. CVB3 has been identified as the most common pathogen for viral myocarditis (VM). CVB3 infections are spread worldwide, and the clinical manifestations are mainly asymptomatic or mild infections, and cold-like symptoms, but newborns and children are more likely to have severe diseases, such as pancreatitis, myocarditis, encephalitis, and type 1 diabetes (Pauschinger et al., 2004; Kühl et al., 2005).

Many factors can affect the infection of the virus, like host protein and non-coding RNAs (ncRNAs). ncRNAs include microRNAs (miRNAs), long non-coding RNA (lncRNA), and circular RNA (circRNA; Salmena et al., 2011). Circular RNAs (circRNAs) are a novel class of ncRNAs, originating from pre-mRNAs. CircRNAs are circularized by connecting a 5′ splice site with the 3′ splice site of an upstream exon or intron by a back-splicing reaction (Wang et al., 2016). Most circRNAs are composed of exons and are located in the cytoplasm; they play a significant role in regulating the translation and modification of proteins (Danan et al., 2012). In some research, it has been shown that circRNAs act as sponges for miRNAs; by forming the competing endogenous RNAs loops, circRNAs could direct binding with specific miRNAs to regulate post-transcriptional gene expression events (Memczak et al., 2013). CircBACH1 regulated hepatitis B virus by miR-200a-3p/MAP 3K2 axis (Du et al., 2022). CircSIAE inhibited CVB3 by targeting miR-331-3p (Yang et al., 2021). CircEAF2 reduced Epstein–Barr virus by miR-BART19-3p/APC/β-catenin axis (Zhao et al., 2021). These researches showed that circRNAs play an important role in infectious diseases.

The hsa_circ_0063331 (circDDX17) was formed by reverse splicing the linear transcript of exons 2–5 of the *DEAD-Box Helicase 17* (*DDX17*) gene with a length of 451 nucleotides. DDX17 was a member of the DEAD-box helicase family proteins involved in cellular RNA folding, splicing, and translation (Linder and Jankowsky, 2011). Moreover, DDX17 was involved in some virus replication, like by binding to specific stem-loop structures of viral RNA to antivirus (Moy et al., 2014). In another study, it could downregulate the expression of Epstein–Barr virus genes by YTH domain-containing proteins recruiting (Xia et al., 2021). Major studies about circDDX17 were focused on cancer, like circDDX17 as a tumor suppressor in colorectal cancer, breast cancer, and colorectal cancer (Li et al., 2018; Lin et al., 2020; Peng and Wen, 2020; Ren et al., 2020), but its function in the virus was still unclear.

N6-methyladenosine (m6A) is intimately associated with three categories of molecular compositions: “writers,” “readers,” and “easers” (Zaccara et al., 2019). Writers are m6A methyltransferases like the methyltransferase-like protein 3 (METTL3) and methyltransferase-like protein 14 (METTL14). Some research showed that METTL3 could regulate virus replication, like METTL3 inhibits Enterovirus 71 by autophagy regulation (Xiao et al., 2021), decreases syndrome coronavirus clade 2 viral load and viral gene expression in host cells (Li et al., 2021), and promotes Epstein–Barr virus infection of nasopharyngeal epithelial cells (Dai et al., 2021). NOTCH1 to 4 are transmembrane receptors that determine cell fate. The NOTCH Receptor 2 (NOTCH2) has been reported to exert distinct functions in regulating tissue homeostasis and cell fate determination (Baron, 2017; Afaloniati et al., 2020). In infectious diseases, NOTCH2 is possibly involved in regulating the Epstein–Barr virus latent/lytic status (Giunco et al., 2015), and 4.3% of hepatitis C virus-positive cells diffuse large B-cell lymphoma have NOTCH2 mutations (Arcaini et al., 2015). Previous studies have shown that NOTCH2 has some relationship with METTL3, like the Notch signaling pathway as an important downstream target of METTL3 in muscle stem cells (Liang et al., 2021), but no data showed that NOTCH2 has a direct relation with METTL3. In this study, we found silence NOTCH2 could downregulate METTL3, and overexpression NOTCH2 could upregulate METTL3. In co-immunoprecipitation analysis, METTL3 was present in the immunoprecipitated complex, and METTL3 was partially co-localized with NOTCH2 in HeLa cells.

Here, we study the effect of CVB3 infection on expression levels of circRNAs and investigate potential downstream mechanisms of their involvement in viral processes *in vivo* and *in vitro*. We examined the prevalence, regulation, and functional roles of circDDX17 in CVB3 infection. CVB3 infection could decrease the expression level of circDDX17 in cells and mics. CircDDX17 up-regulated CVB3 replication in HeLa and HL-1 cells. Furthermore, CircDDX17 regulated NOTCH2 by target miR-1,248, miR-1,248 could down-regulate NOTCH2 expression and inhibit CVB3 replication.

# Materials and methods

## Cells and virus

HeLa and HEK-293T cells were a gift from Dr. Huaiqi Jing (Chinese Center for Disease Control and Prevention). HL-1 cells were stored at the School of Medicine, Jiangsu University. Cells were cultured with Dulbecco’s modified Eagle’s medium (DMEM, Gibco, United States), supplemented with 8% fetal bovine serum (FBS, Gibco, United States), 100 U/ml penicillin, and 100 μg/ml streptomycin in 5% CO$^2$ at 37°C. CVB3 (Nancy; Corsten et al., 2015) was a gift from Professor Ruizhen Chen (Department of Cardiology, Zhongshan Hospital, Shanghai, China). GFP-CVB3, expressing the green fluorescence protein (GFP; Lei et al., 2013; Shuo et al., 2014).

## Myocarditis

Coxsackievirus B3 ($10^5$ PFU/mouse) was injected intraperitoneally into 3-week-old BALB/c male mice. CVB3 was diluted in 100 μl PBS for injection, and an equal volume of PBS was injected into the blank control mice (3 mice per group). This study was conducted according to the recommendations in the Guide to the Care and Use of Experimental Animals-Chinese Council on Animal Care. All protocols were approved by the Animal Care Committee of University Jiangsu (protocol number: UJS-IACUC-AP-20190307087).

## Plasmid, miRNA, siRNA, and transfection

PcicR-3.0-circDDX17 (pcircDDX17) for overexpression circDDX17, PcicR-3.0 (pcicR) for its negative control. PcDNA-3.0-NOTCH2 (pNOTCH2) for overexpression NOTCH2, used pcDNA-3.0 (pcDNA) for its negative control. miR-885 mimics (miR-885), miR-1248 mimics (miR-1248), and miR-1279 mimics (miR-1279), negative control (miR-NC); miR-1,248-inhibitor (miR-1,248-in), negative control (NC-in); siRNA-NOTCH2 (si-NOTCH2), negative control (si-NOTCH2-NC) were all synthesized by GenePharma Co., Ltd. (Suzhou, China), the sequences were listed in Table 1. Plasmid and oligonucleotide were transfected using Lipofectamine 3000™ (Invitrogen, United States).

## RNA preparation and quantitative real-time PCR

The total RNA was isolated using Trizol reagent (Invitrogen, United States). PrimeScript RT Reagent Kit (Takara, Japan) was used for reverse transcription RNA, and quantitative real-time PCR (RT-qPCR) was performed using TB Green Premix Ex TaqII (Takara, Japan). The RT-qPCR was conducted to examine the expression levels of circDDX17, mRNA levels for GAPDH, NOTCH2, and VP1, and miRNA levels for miR-885, miR-1248, miR-1279, and U6. The divergent primer was synthesized by Sangon (China), and the sequences are listed in Table 2. For RNase treatment, 2 mg of total RNA was incubated with or without 3 U/mg RNase R for 30 min at 37°C.

## Cytoplasmic nucleus separation

According to the manufacturer's instructions, nuclear plasma was extracted with a cytoplasmic nucleus extraction kit (Thermo Fisher, United States); the steps were followed as previously described (Yang et al., 2021).

TABLE 1 Sequence details.

| Name | Sequence |
|---|---|
| si-circDDX17 | 5′ GGCCCAAUCAUUUGGAGCATT 3′ |
| miR-1,248 mimics | 5′ ACCUUCUUGUAUAAGCACUGUGCUAAA 3′ |
| miR-885-5p mimics | 5′ UCCAUUACACUACCCUGCCUCU 3′ |
| miR-1,279 mimics | 5′ UCAUAUUGCUUCUUUCU 3′ |
| miR-1,248 inhibitor | 5′ UUUAGCACAGUGCUUAUACAAGAAGGU′ |
| si-NOTCH2-1 | 5′ GGCAGUGUGUGGAUAAAGUTT 3′ |
| si-NOTCH2-2 | 5′ GGAGGUCUCAGUGGAUAUATT 3′ |
| si-NOTCH2-3 | 5′ GUGCCAGACAGACAUGAAUTT 3′ |

## Fluorescence *in situ* hybridization (FISH)

Cy5-labeled circDDX17 probes (Jima Biotech, China) were detected in HeLa cells using a Fluorescent *in Situ* Hybridization Kit (Jima Biotech, China) following the manufacturer's guidelines. Cell nuclei were counterstained with DAPI (Jima Biotech, China). The glass slides were analyzed and images were captured under a fluorescence microscope.

## Immunofluorescence (IF) microscopy

Cells cultured in collagen-coated chamber slides (NEST Biotechnology Co., Ltd., China) were washed and fixed with either 4% paraformaldehyde or with ice-cold methanol. Cells were permeabilized with 0.1% Triton X-100 in PBS, slides were stained with all primary antibodies (anti-METTL3, 1:200, anti-NOTCH2, 1:100), washed three times with PBS, and stained with conjugated Alexa Fluor secondary antibodies Alexa Fluor 488/594 (1200, Genetex, United States), cell nuclei were counterstained with DAPI (Jima Biotech, China). The glass slides were analyzed and images were captured under a fluorescence microscope.

## Western blot

The total proteins of the cells were extracted using the RIPA lysis buffer (Sigma, United States). Samples were subjected to 12% SDS-PAGE and transferred to polyvinylidene fluoride (PVDF) membranes (Millipore, United States). The primary antibodies against NOTCH2 (1:1,000, Sangon, China), VP1 (1:1,000, Genetex, United States), METTL3 (1:2,000, CST, United States), methyltransferase-like protein 14 (METTL14; 1:2,000, CST, United States), and GAPDH (1:20000, Genetex, United States). Membranes were blocked, incubated with secondary antibody (1:20000, Jackson, United States), and detected by electrochemiluminescence (ECL, Millipore, United States).

## Co-immunoprecipitation

For coimmunoprecipitation (co-IP) assays, 500 μg HeLa cell lysate protein was reacted with primary antibodies (10 μl) overnight at 4°C and incubated with protein A/G beads at the next day for 4 h at 4°C. Then immunoprecipitated proteins were eluted from the beads for Western blotted with indicated antibodies.

## Luciferase reporter assays

HEK-293 T cells ($1 \times 10^5$/well) were plated in 24-well plates. Cells were co-transfected with miR-1248 and psiCHECK-2 luciferase reporter to generate psiCHECK-2-circDDX17-wild-type (circDDX17-wt) or psiCHECK-2-circDDX17-mutant

TABLE 2 Primer sequence details.

| Name | Forward Primer (5′–3′) | Reverse Primer (5′–3′) |
|---|---|---|
| hsa_circDDX17 | ATTTCCGTTGGCTCTTAGTG | CCTCTTGCTCCAAATGATTG |
| hsa-GAPDH | AGGTGAAGGTCGGAGTCAAC | GGGTGGAATCATATTGGAACA |
| mus_circDDX17 | ATTTCCTTTGGCTCTTAGTG | CCAGCACTAGACAAATTC |
| mus-GAPDH | TGCCCCCATGTTTGTGATG | TGTGGTCATGAGCCCTTCC |
| VP1 | ATTCAAGGTCCGAGTCAAC | CTGCTTGTCGTGGTGTTA |
| hsa-NOTCH2 | CGGGGCCTACTCTGTGAAGA | ACTACGGCAAACACACAGGT |
| U6 | CTCGCTTCGGCAGCACA | AACGCTTCACGAATTTGCGT |
| hsa-miR-885-5p-mimics | CTCAACTGGTGTCGTGGAGTCGGCAATTCAGTTGAGAGAGGCAG | ACACTCCAGCTGGGTCCATTACACTACC |
| hsa-miR-1248 mimics | CTCAACTGGTGTCGTGGAGTCGGCAATTCAGTTGAGTTTAGCAC | ACACTCCAGCTGGGACCTTCTTGTATAAGCACTG |
| hsa-miR-1279 mimics | CTCAACTGGTGTCGTGGAGTCGGCAATTCAGTTGAGAGAAAGAA | ACACTCCAGCTGGGTCATATTGCTT |

(circDDX17-mu) constructs after 48h. Luciferase activity was determined following a dual-luciferase reporter assay detection kit (Promega, Madison, WI, United States).

## Plaque assay

Sample supernatants were collected at CVB3 7h post-infection, serially diluted, and added onto HeLa cells in 24-well plates ($1.0 \times 10^5$ cells/well). After incubation for 1h, cells were rewashed with PBS three times, overlaid with 0.75% soft agar medium, and incubated for 3 days. Cells were fixed with glacial acetic acid for 30min and stained with 1% crystal violet. All the assays were conducted at least in triplicate.

## Statistical analysis

Statistical analysis was performed using Graph Pad Prism 7 (GraphPad Software Inc., San Diego, CA, United States). The group difference was evaluated by Student's *t*-test, error bars represent mean ± SD and the difference was statistically significant when $*p < 0.05$, $**p < 0.01$, or $***p < 0.001$.

# Results

## Result 1. Characterization of the existence and subcellular distribution of circDDX17 in HeLa cells

The Circbank database showed that circDDX17 is formed by reverse splicing of the linear transcript of exons 2–5 of the *DDX17* gene with a length of 451 nucleotides. To further characterize circDDX17, Sanger sequencing was performed to confirm head-to-tail splicing (Figures 1A,B). FISH analysis (Figure 1C) and nuclear separation assay (Figures 1D,E) was conducted to determine the subcellular localization of circDDX17 in HeLa cell lines.

To investigate the roles of circDDX17 in CVB3 infection, HeLa and HL-1 cells were infected with CVB3, and circDDX17 levels were examined. In HeLa cells, at the early stages of infection (0–2h), no significant change in circDDX17 expression was detected between CVB3-infected and mock control cells. At 4–7h post-infection, the circDDX17 expression was significantly decreased compared to MOCK cells (Figure 1F). The expression level of circDDX17 in HL-1 cells infected with 200 MOI CVB3 was lower than 100 MOI CVB3 at 24h post-infection (Figure 1G). In VM mice, the circDDX17 expression level was lower than in MOCK mice (Figure 1H).

## Results 2. CVB3 infection reduced circDDX17 expression level, and circDDX17 promotes CVB3 replication

HeLa cells (Figures 2A,B) and HL-1 cells (Figure 2C) overexpressed circDDX17 were infected with CVB3 to analyze the expression of VP1. The results showed that circDDX17 increased the expression of VP1 after CVB3 infection. HEK-293T cells co-transfected with pcicR-3.0-circDDX17 (pcircDDX17) and GFP-CVB3, cells overexpressed circDDX17 GFP-positive cell number were observed more than cells co-transfected with pcicR-3.0 (pcicR) and GFP-CVB3 (Figure 2D). To gain further insight into the function of cricDDX17 on CVB3 replication, a viral plaque assay was adopted. The results showed that cricDDX17 overexpression increased viral release compared to the control (Figure 2E).

In contrast, circDDX17 silencing decreased the expression of VP1 at CVB3 after CVB3 infection (Figures 2F–H). HEK-293T cells co-transfected with siRNA-circDDX17 (si-circDDX17) and GFP-CVB3 showed lower GFP-positive cells than the cells

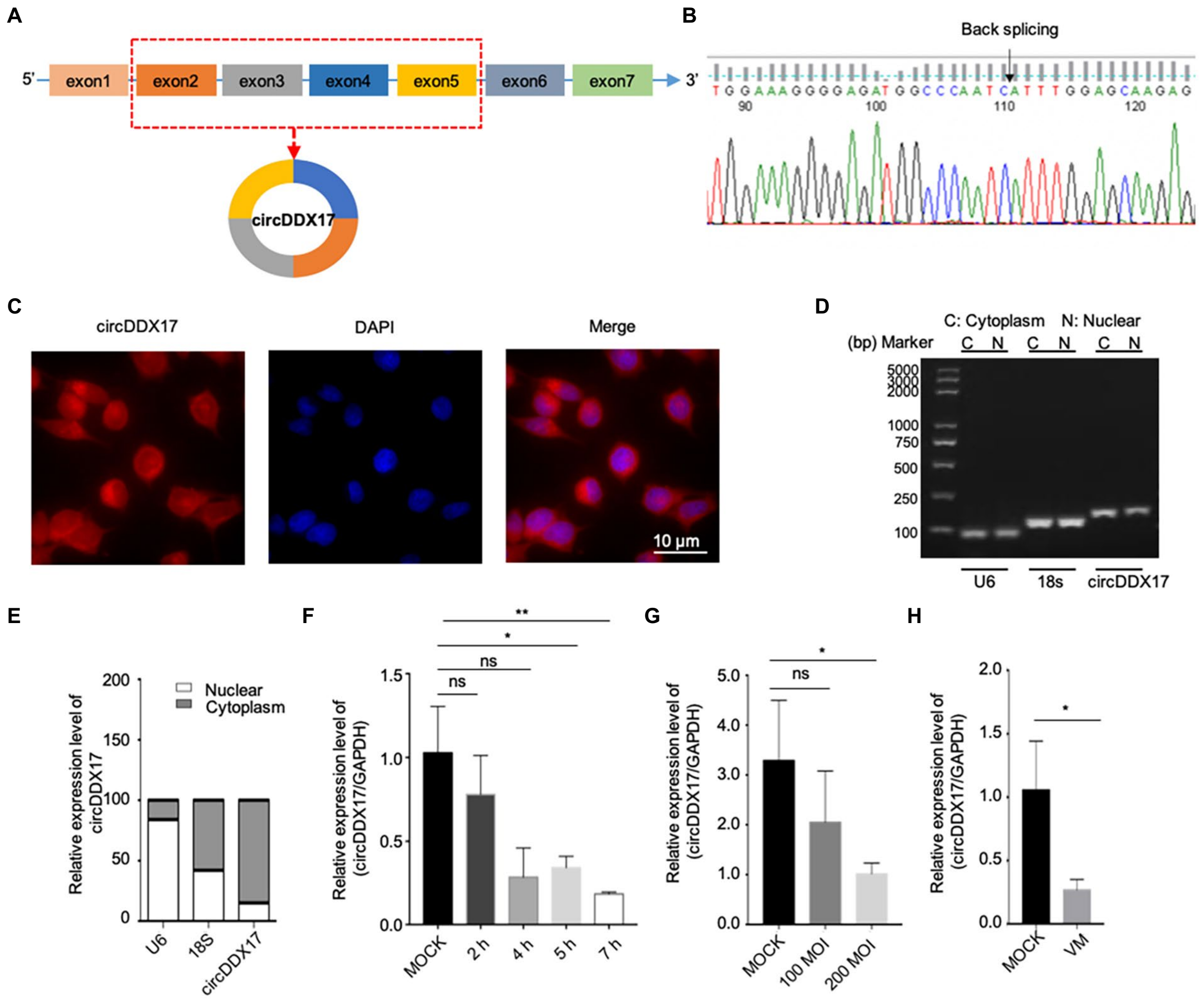


FIGURE 1
Characterization and subcellular distribution of circDDX17 in HeLa cells. **(A,B)**. Sanger sequencing confirmation of the head-to-tail splicing of circDDX17. **(C–E)** The sub-cellular distribution of circDDX17 was mostly present in the cytoplasm by the nuclear mass. **(F)** RT-qPCR analysis of circDDX17 expression in HeLa cells 7h after CVB3 infection. **(G)** RT-qPCR analysis of circDDX17 in CVB3-infected HL-1 cells (100 MOI and 200 MOI) for 24h. **(H)** RT-qPCR analysis of circDDX17 in viral myocarditis, MOCK mice were as control. *$p$<0.05, **$p$<0.01.

transfected siRNA-circDDX17-NC (si-circDDX17-NC) and GFP-CVB3 (Figure 2I), and circDDX17 silencing could reduce viral release (Figure 2J).

## Results 3. CircDDX17 promotes CVB3 replication *via* the miR-1248/NOTCH2 axis

CircRNAs have been shown to act as a miRNA sponge to regulate gene expression (Guo et al., 2020); therefore, the potential miRNAs associated with circDDX17 were investigated. First, through bioinformatics analysis, three miRNAs (miR-885, miR-1248, and miR-1279) were identified from the overlap of three databases (circBank, miRanda, and CircInteractome) as possible targets for circDDX17. RT-qPCR of HeLa cells transfected with pcircDDX17 or si-circDDX17 showed that circDDX17 downregulates the miRNAs expression (Figures 3A–C). To analyze the role of CVB3 on miRNA expression, total RNA from CVB3 infected cells was collected for RT-qPCR. miR-885, miR-1248, and miR-1279 expression increase with virus infection (Figures 3D–F). To investigate the role of miRNAs in CVB3 infection, HeLa cells were transfected with miRNAs mimics, and the results showed that miR-885, miR-1248, and miR-1279 could inhibit the replication of CVB3, among which miR-1248 had the most obvious effect (Figure 3G), so we chose miR-1248 as the object of study. The dual-luciferase reporter assay confirmed the direct interaction between circDDX17and miR-1248. The circDDX17-wild-type (circDDX17-wt) and circDDX17-mutant (circDDX17-mu) full-length sequences without miR-1248 binding sites were cloned into the luciferase vector. Subsequently, luciferase reporter assays confirmed that miR-1248 mimics markedly

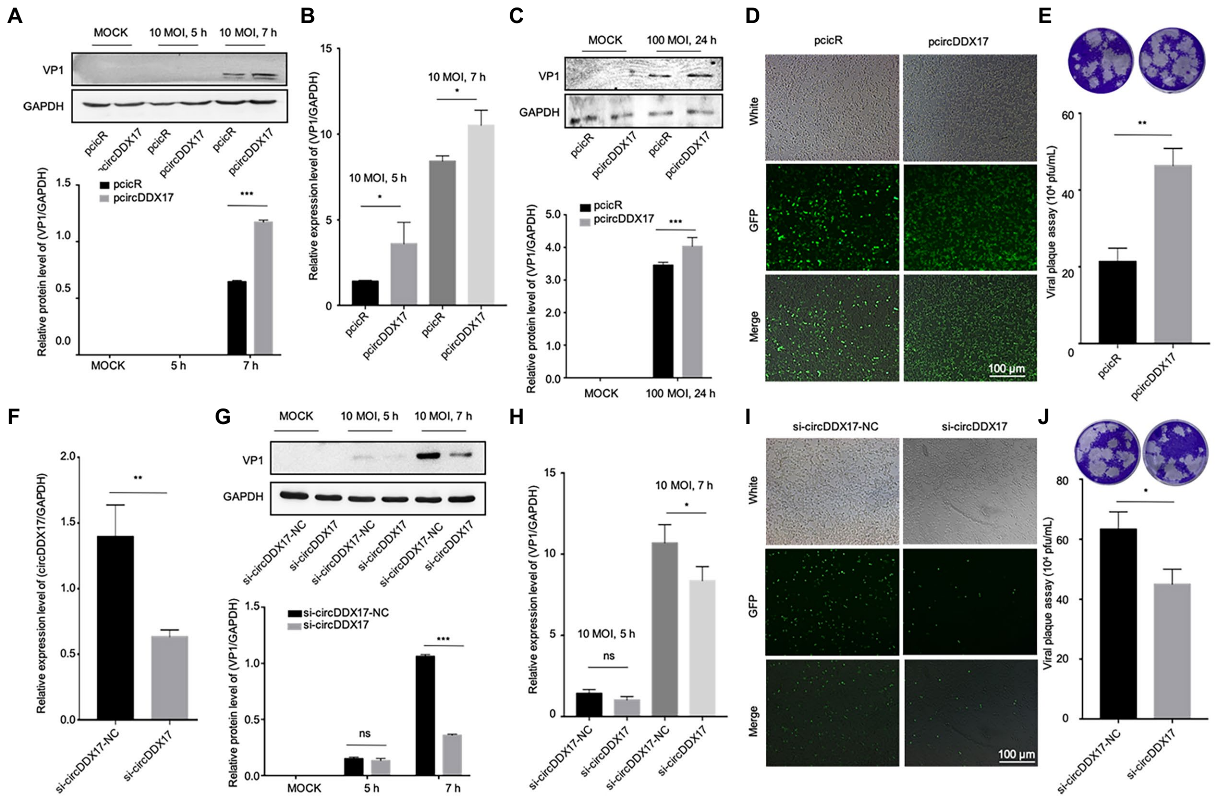


FIGURE 2
CircDDX17 promotes CVB3 replication. **(A,B)** Western blot and RT-qPCR of VP1 expression levels in CVB3-infected HeLa cells overexpressing circDDX17 (Western blot and analyzed using ImageJ software). **(C)** Western blot to analyze the expression of VP1 after HL-1 overexpression circDDX17. **(D)** HEK-293T cells co-transfected pcircDDX17 and GFP-CVB3, observed GFP-positive cell number after 48h. **(E)** Viral plaque assay. Viral titer was determined by plaque assay using the supernatants collected at 7h (*n*=3). **(F)** RT-qPCR to analyze the circDDX17 expression after si-circDDX17 transfected HeLa cells. **(G,H)** Western blot and RT-qPCR to detect VP1 expression in HeLa cells after transfected with the si-circDDX17 and infected CVB3 (10 MOI). **(I)** HEK-293T co-transfected with si-circDDX17 and GFP-CVB3, observed GFP-positive cell number after 48h. **(J)** Viral titers were determined by plaque assay using the supernatants (*n*=3). **p*<0.05, ***p*<0.01, ****p*<0.001.

reduced the luciferase activity of circDDX17-wt but not that of circDDX17-mu compared to the miR-NC group (Figure 3H). To investigate the signal pathways contributing to the miR-1,248 effect on CVB3 replication, we sought to identify its target genes. Bioinformatic analyses using TargetScan, miRDB, and miRWalk programs showed that NOTCH2 is one of the predicted targets. HeLa cells were transfected with miR-1248 mimics (miR-1248) or miR-1248-inhibitor (miR-1248-in), and the results showed that miR-1248 has a negative regulatory role on NOTCH2 expression (Figure 3I).

In previous research, NOTCH2 has a relationship between METTL3 and DNA-methylation (Terragni et al., 2014), we predicted METTL3 as NOTCH2 interaction protein by String,[1] therefore, HeLa cells were transfected with specific si-NOTCH2 or pNOTCH2 to silence or overexpressed NOTCH2. For further verification, we performed coimmunoprecipitation experiments to study the relationship between NOTCH2 and METTL3 in HeLa cells (Figure 3J), the result showed that the METTL3 was present in the immunoprecipitated complex. As shown in Figure 3K, METTL3 and NOTCH2 were distributed in the nucleus although a small fraction of these proteins were also found in the cytoplasm, and METTL3 was partially co-localized with NOTCH2. The Western blot results show that NOTCH2 regulates METTL3 expression, METTL3 increases with the increasing expression level of NOTCH2 and decreases with NOTCH2 decreasing expression level (Figures 3L,M).

## Results 4. CircDDX17 promotes CVB3 replication, and rise DNA-methylation-associated protein METTL3 and METTL14 expression

To elucidate the mechanism of CVB3 upregulation of host circDDX17, the expressions of DNA-methylation-associated proteins METTL3, and METTL14 were examined. With CVB3

[1] https://cn.string-db.org/

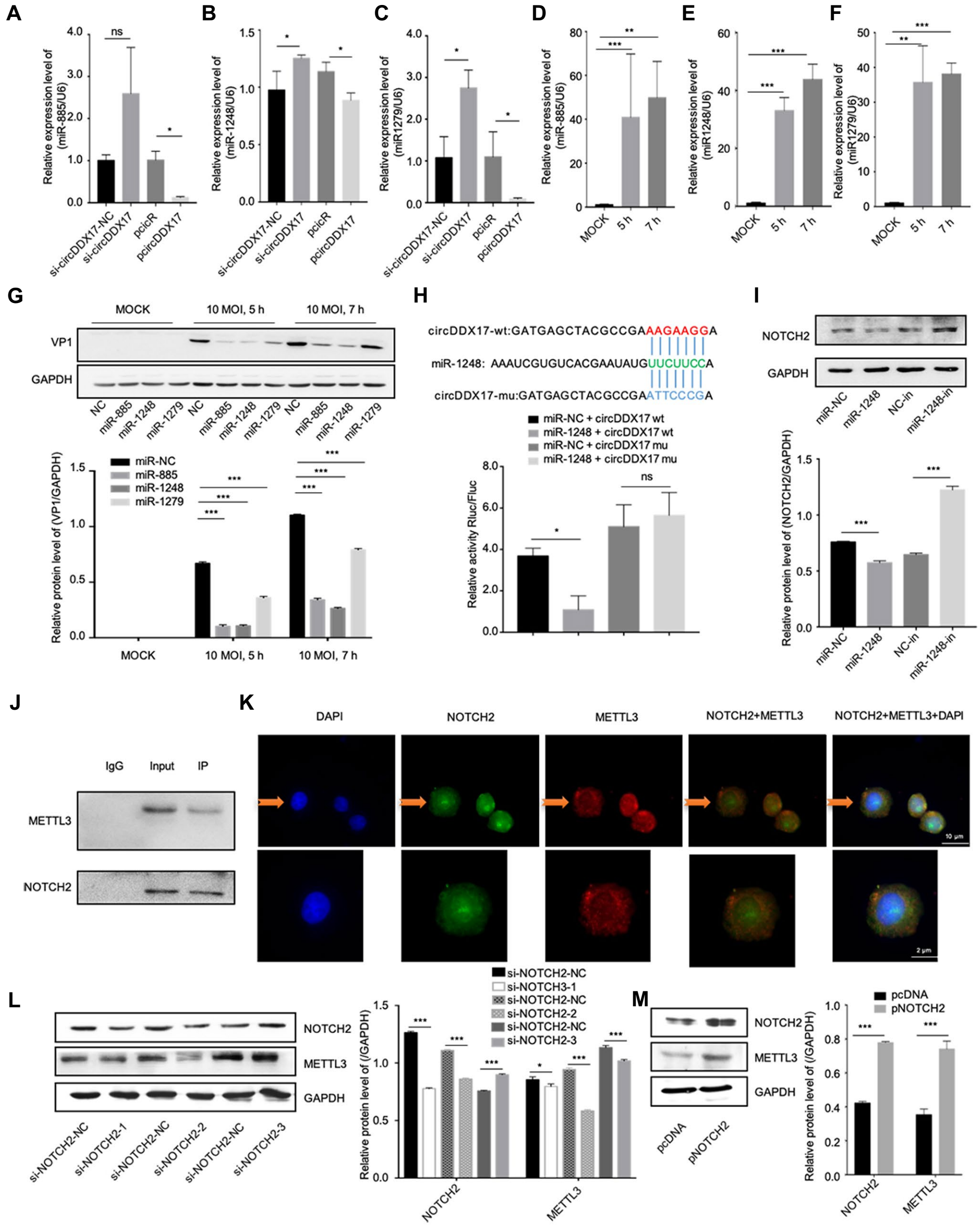


FIGURE 3

CircDDX17 promotes CVB3 replication by targeting miR-1248, and miR-1248 targeting NOTCH2. **(A–C)** Relative expression levels of miR-885, miR-1248, and miR-1279 in HeLa cells were determined by RT-qPCR after cells overexpression or silence the circDDX17. **(D–F)** RT-qPCR to detected expression of miR-885, miR-1248, and miR-1279 in CVB3 (10 MOI) infected HeLa cells. **(G)** HeLa cells transfected with miR-885, miR-1248, and miR-1279, infected the CVB3 (10 MOI) for 7h. Western blot analysis of the VP1 expression. **(H)** The putative miR-1248 binding site in circDDX17 (circDDX17-wt) and the designated mutant sequence (circDDX17-mu) are illustrated. Validation of circDDX17 targeting on miR-1248 luciferase assay. **(I)** HeLa cells transfected miR-1248 mimics or miR-1248-in, indicated NOTCH2 detected by Western blot.
**(J)** Immunoprecipitation and immunoblot analyses were performed with the indicated antibodies. **(K)** HeLa cells stained with anti-NOTCH2 (green) and anti-METTL3 (red) antibody and were analyzed by confocal microscope. Nuclei were labeled with DAPI (blue). Micrographs with ×40 magnification (scale bar of 10μm) are shown. **(L)** HeLa cells transfected si-NOTCH2, indicated signals were detected by Western blot. **(M)** NOTCH2 overexpression in HeLa cells. Indicated signals were detected by Western blot. *$p$<0.05, **$p$<0.01, ***$p$<0.001.

infection, NOTCH2 declined gradually, and METTL3 and METTL14 were elevated (Figures 4A–D). To understand the roles of NOTCH2 in circDDX17-mediated DNA-methylation, we examined the downstream effector gene expression after CVB3 infection in HeLa cells overexpressing or silencing circDDX17. Western blot analysis showed that overexpression of circDDX17 upregulated the expression level of NOTCH2, METTL3, and METTL14 (Figures 4E,F). Conversely, circDDX17

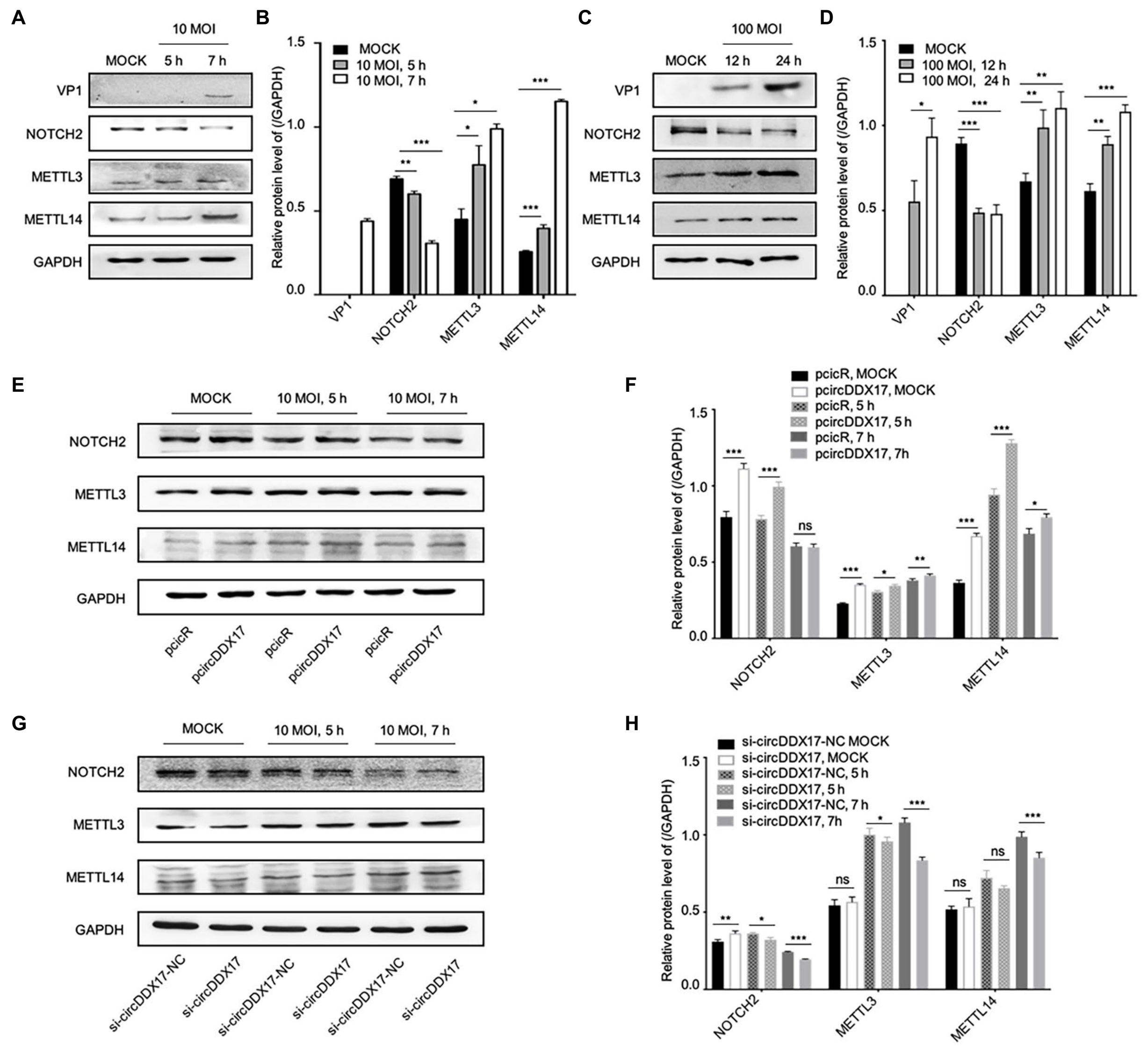


FIGURE 4

CircDDX17 regulates NOTCH2 expression and influences methylation-related pathways after CVB3 infection. **(A,B)** HeLa cells infected CVB3 (10 MOI) for 7h, signals were detected by Western blot. **(C,D)** HL-1 cells infected CVB3 (100 MOI) for 24h, signals were detected by Western blot. **(E,F)** HeLa cells overexpression circDDX17 by pcircDDX17, cells infected CVB3 (10 MOI) for 7h, indicated signals were detected by Western blot. **(G,H)** HeLa cells silenced circDDX17 by si-circDDX17, cells infected CVB3 (10 MOI) for 7h, indicated signals were detected by Western blot. **p*<0.05, ***p*<0.01, ****p*<0.001.

silencing decreases NOTCH2, METTL3, and METTL14 expression levels (Figures 4G,H).

## Results 5. CircDDX17 promotes CVB3 replication by targeting miR-1248.

To study the effect of miR-1,248 in NOTCH2, METTL3, and METTL14 expression, HeLa cells were transfected with miRNA mimics or miRNA inhibitors, Western blot showed that miR-1248 played a negative role on NOTCH2 expression, and so did METTL3 and METTL14 expression levels (Figure 5A). Then cells infected with CVB3, miR-1248 decreased VP1 expression while inhibiting miR-1248 increased VP1 expression, and miR-1248 down-regulate the NOTCH2, METTL3, and METTL14 expression (Figure 5B).

HeLa cells overexpressing miR-1248 reduced CVB3 replication and cells silencing miR-1248 increased CVB3 replication (Figures 5C,D). At the same time, miR-1248 decreased CVB3 released as detected by viral plaque assay (Figure 5E).

To further confirm that miR-1248 downregulation by circDDX17 benefits CVB3 replication, we overexpressed miR-1248 in the presence of circDDX17 by co-transfection, and cells without CVB3 (Figure 5F) or infected CVB3 (Figure 5G).

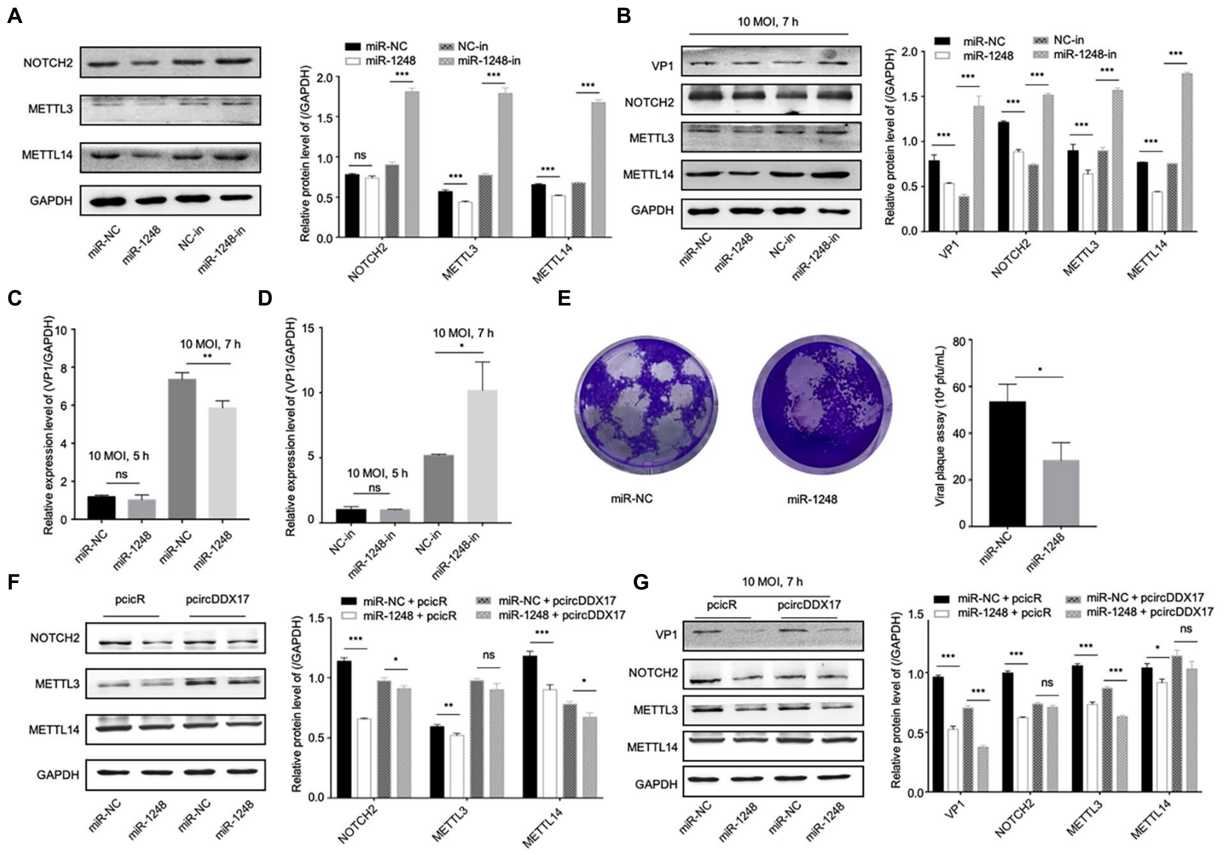


FIGURE 5
CircDDX17 regulates NOTCH2 expression by targeting miR-1248, influence the methylation-related pathway after CVB3 infection. **(A)** HeLa cells transfected miR-1248 mimics or miR-1248-inhibitor, indicated signals were detected by Western blot. **(B)** HeLa cells transfected miR-1248 mimics or miR-1248-inhibitor, and infected CVB3 (10 MOI) for 7h indicated signals were detected by Western blot. **(C)** HeLa cells transfected the miR-1248, infected CVB3 (10 MOI) for 7h, RT-qPCR to detect the VP1 expression. **(D)** HeLa cells silenced miR-1248 and infected CVB3 (10 MOI) for 7h, RT-qPCR to detect the VP1 expression. **(E)** Viral titers were determined by plaque assay using the supernatants of HeLa cells overexpression miR-1,248 (*n*=3). **(F)** HeLa cells co-transfected miR-1248 and pcircDDX17, indicating signals were detected by Western blot. **(G)** HeLa cells co-transfected miR-1248 and pcircDDX17, cells infected CVB3 (10 MOI) for 7h, indicated signals were detected by Western blot. **p*<0.05, ***p*<0.01, ****p*<0.001.

Western blot showed that compared with miR-NC + PcicR, overexpression of miR-1248 inhibited VP1, NOTCH2, METTL3, and METTL14 expression.

## Results 6. MiR-1248 inhibits CVB3 replication by targeting NOTCH2

To understand the roles of NOTCH2 in circDDX17 and miR-1248-mediated methylation-related pathways, we first confirmed NOTCH2 function on methylation-related proteins. Western blot data showed that independent of CVB3 infection, METTL3, and METTL14 expression was reduced by si-NOTCH2 transfection and induced in pNOTCH2 transfection (Figure 6A). In HeLa cells, VP1 expression was decreased by si-NOTCH2 and increased by pNOTCH2 expression (Figure 6B). Furthermore, in HL-1 cells, overexpression of NOTCH2 increased VP1 expression compared to the negative control (Figure 6C). NOTCH2 could significantly increase VP1 expression level, NOTCH2 deficiency repressed VP1 expression (Figures 6D,E), and overexpression NOTCH2 increased viral release (Figure 6F).

To further confirm miR-1248 regulated CVB3 replication by targeting NOTCH2, HeLa cells overexpression miR-1248 and NOTCH2 by co-transfection (Figures 6G,H). Without CVB3 infection, cells transfection miR-NC + pNOTCH2 the METTL3 and METTL14 expression levels were higher than cells transfection miR-1248 + pNOTCH2 (Figure 6G). Seven hours post CVB3 infection, miR-NC + pNOTCH2 increased the production of VP1, METTL3, and METTL14 compared with miR-NC + pcDNA. Co-transfection of pNOTCH2 and miR-1248 reduced the production of VP1, METTL3, and METTL14 compared to miR-NC + pNOTCH2 (Figure 6H).

## Discussion

Coxsackievirus B3 is the commonest pathogen for acute and chronic myocarditis (Pauschinger et al., 2004; Esfandiarei and McManus, 2008). After CVB3 entry into the cardiomyocytes, the

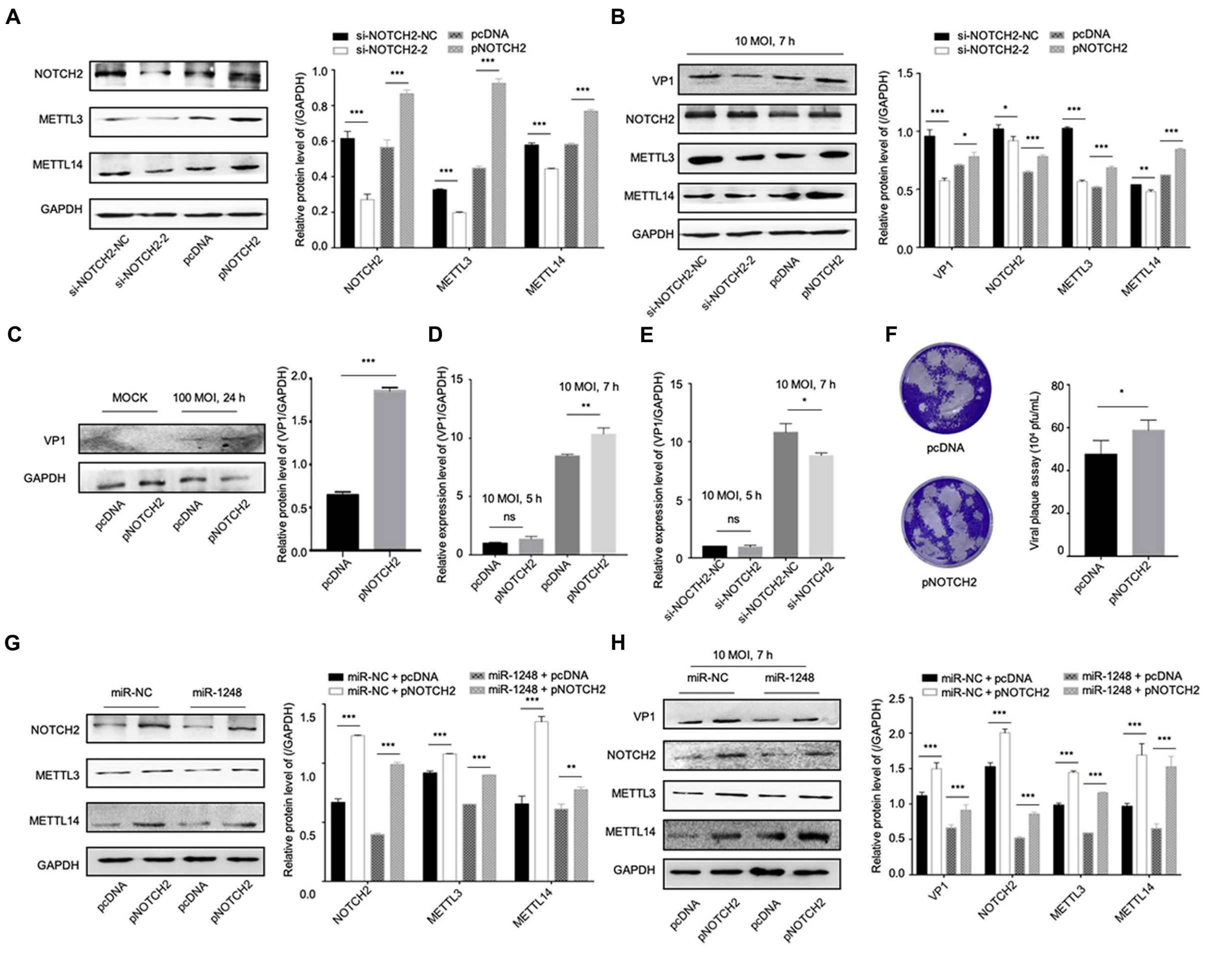


FIGURE 6
MiR-1248 regulates CVB3 replication by targeting NOTCH2 through methylation-related pathways. **(A)** HeLa cells transfected with pNOTCH2 or pcDNA, detected by Western blot. **(B)** HeLa cells transfected with pNOTCH2 or pcDNA and infected CVB3 (10 MOI) for 7h, detected by Western blot. **(C)** HL-1 cells transfected the pNOTCH2, infected CVB3 (100 MOI) for 24h, and Western blot analyzed the VP1. **(D,E)** HeLa cells transfected pNOTCH2 or si-NOTCH2-2, infected CVB3 (10 MOI) for 7h, and RT-qPCR to detect the VP1 expression. **(F)** Viral titers were determined by plaque assay using the supernatants ($n$=3). **(G)** HeLa cells co-transfected miR-1248 and pNOTCH2, Western blot to analyze the signals. **(H)** HeLa cells co-transfected miR-1248 and pNOTCH2, cells infected CVB3 (10 MOI) for 7h, indicated signals were detected by Western blot. *$p$<0.05, **$p$<0.01, ***$p$<0.001.

virus replicates and induces cell damage, triggering the host immune responses. If the virus cannot be eliminated, myocarditis can become chronic, triggering extensive myocardial fibrosis and the development of dilated cardiomyopathy (Kawai, 1999; Garmaroudi et al., 2015). In our previous study, miR-324-3p inhibits CVB3 replication by targeting the tripartite motif 27 (Liu et al., 2021), but there are fewer studies on circRNA regulation of CVB3 replication. CircRNA can play a role as a miRNA sponge to influence miRNA expression and regulate gene function.

This study identified that circDDX17 was a novel regulator of CVB3 replication. MiR-1248, a target miRNA of circDDX17, played a negative role in replicating CVB3 in host cells. Moreover, NOTCH2 was the miR-1248 target gene. NOTCH2 has been involved in cardiac fibrosis, regulating heart development and multiple antiviral immune responses. In addition, NOTCH2 mutations resulted in multiple cardiac diseases and vascular anomalies (Pinkert et al., 2019). Interestingly, NOTCH2 was distributed in m6A modification proteins. m6A was a conserved internal modification found in almost all eukaryotic nuclear RNAs (Jia et al., 2013) and the viral RNA. m6A was dynamic methylation involved in RNA metabolism, splicing, and decay (Roundtree et al., 2017; Zhao et al., 2017). METTL3 could modulate the NOTCH signaling pathway (Wang et al., 2020). In our study, there was a positive correlation between NOTCH2 and METTL3. By the analysis of IP, METTL3 was present in the immunoprecipitated complex, and METTL3 was partially co-localized with NOTCH2, those results showed that METTL3 and NOTCH2 have interaction in cells. METTL3 could negatively regulate type I interferon response by dictating the fast turnover of interferon mRNAs for antivirus (Winkler et al., 2019), and METTL3 boosted Enterovirus 71 replication (Hao et al., 2019), which might explain how NOTCH2 regulates viral replication. In another way, m6A modification was dynamically and reversibly regulated by the "writers" complex (METTL3 and METTL14; Liu et al., 2014). Our study analyzed METTL3 and METTL14 as targets indicating m6A modification changes and function in cells overexpressing or silencing circDDX17 infected with CVB3. The results showed that CVB3 infection could increase the METTL3 and METTL14 expression. METTL14 played an important role in the transcription of IFNs and inflammatory cytokines, and regulates antivirus innate immunology response (Xu et al., 2021). However, in this research, we have not studied the effect of METTL14 on the replication of CVB3.

In conclusion, this study reported that circDDX17 promotes CVB3 replication by regulating miR-1248 and NOTCH2/METTL3. These findings enriched our understanding of the functional roles of circRNA in viral replication and provided novel insights into the development of therapeutic strategies.

## Data availability statement

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

## Ethics statement

The animal study was reviewed and approved by Animal Care Committee of University Jiangsu (protocol number: UJS-IACUC -AP-20190307087).

## Author contributions

HS and HW conceived and designed the experiments. TL, YL, XL, QY, YW, and XQ performed the experiments. HS, HW, and SS analyzed the data. TL, HS, HW, JT, XD, and SC contributed reagents, materials, and analysis tools. TL, HS, and HW wrote the paper. All authors contributed to the article and approved the submitted version.

## Funding

The research was supported by National Natural Science Foundation of China, grant no. 81971945 (https://isisn.nsfc.gov.cn/egrantweb/).

## Acknowledgments

We thank Songfeng Zhou (Nantong University, Nantong China) for providing the pcicR-3.0. We are grateful to Zhaohua Zhong (Harbin Medical University, Harbin, China), who provided GFP-CVB3.

## 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.2022.1012124/full#supplementary-material

## References

Afaloniati, H., Karagiannis, G., Karavanis, E., Psarra, T., Karampatzakis-Kouritas, A., Poutahidis, T., et al. (2020). Inflammation-induced colon cancer in uPA-deficient mice is associated with a deregulated expression of notch signaling pathway components. *Mol. Cell. Biochem.* 464, 181–191. doi: 10.1007/s11010-019-03659-9

Arcaini, L., Rossi, D., Lucioni, M., Nicola, M., Bruscaggin, A., Fiaccadori, V., et al. (2015). The NOTCH pathway is recurrently mutated in diffuse large B-cell lymphoma associated with hepatitis C virus infection. *Haematologica* 100, 246–252. doi: 10.3324/haematol.2014.116855

Baron, M. (2017). Combining genetic and biophysical approaches to probe the structure and function relationships of the notch receptor. *Mol. Membr. Biol.* 34, 33–49. doi: 10.1080/09687688.2018.1503742

Corsten, M., Heggermont, W., Papageorgiou, A. P., Deckx, S., Tijsma, A., Verhesen, W., et al. (2015). The micro RNA-221/−222 cluster balances the antiviral and inflammatory response in viral myocarditis. *Eur. Heart J.* 36, 2909–2919. doi: 10.1093/eurheartj/ehv321

Dai, D., Li, X., Wang, L., Xie, C., Jin, Y., Zeng, M., et al. (2021). Identification of an N6-methyladenosine-mediated positive feedback loop that promotes Epstein-Barr virus infection. *J. Biol. Chem.* 296:100547. doi: 10.1016/j.jbc.2021.100547

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in archaea. *Nucleic Acids Res.* 40, 3131–3142. doi: 10.1093/nar/gkr1009

Du, N., Li, K., Wang, Y., Song, B., Zhou, X., and Duan, S. (2022). Circ RNA circ BACH1 facilitates hepatitis B virus replication and hepatoma development by regulating the mi R-200a-3p/MAP 3K2 axis. *Histol. Histopathol.* 30:18452. doi: 10.14670/HH-18-452

Esfandiarei, M., and Mcmanus, B. (2008). Molecular biology and pathogenesis of viral myocarditis. *Annu. Rev. Pathol.* 3, 127–155. doi: 10.1146/annurev.pathmechdis.3.121806.151534

Garmaroudi, F., Marchant, D., Hendry, R., Luo, H., Yang, D., Ye, X., et al. (2015). Coxsackievirus B3 replication and pathogenesis. *Future Microbiol.* 10, 629–653. doi: 10.2217/fmb.15.5

Giunco, S., Celeghin, A., Gianesin, K., Dolcetti, R., Indraccolo, S., and De Rossi, A. (2015). Cross talk between EBV and telomerase: the role of TERT and NOTCH2 in the switch of latent/lytic cycle of the virus. *Cell Death Dis.* 6:e1774. doi: 10.1038/cddis.2015.145

Guo, X., Dai, X., Liu, J., Cheng, A., Qin, C., and Wang, Z. (2020). Circular RNA circREPS2 acts as a sponge of miR-558 to suppress gastric cancer progression by regulating RUNX3/β-catenin signaling. *Mol. Ther. Nucleic Acids* 21, 577–591. doi: 10.1016/j.omtn.2020.06.026

Hao, H., Hao, S., Chen, H., Chen, Z., Zhang, Y., Wang, J., et al. (2019). N6-methyladenosine modification and METTL3 modulate enterovirus 71 replication. *Nucleic Acids Res.* 47, 362–374. doi: 10.1093/nar/gky1007

Jia, G., Fu, Y., and He, C. (2013). Reversible RNA adenosine methylation in biological regulation. *Trends Genet.* 29, 108–115. doi: 10.1016/j.tig.2012.11.003

Kawai, C. (1999). From myocarditis to cardiomyopathy: mechanisms of inflammation and cell death: learning from the past for the future. *Circulation* 99, 1091–1100. doi: 10.1161/01.cir.99.8.1091

Kühl, U., Pauschinger, M., Seeberg, B., Lassner, D., Noutsias, M., Poller, W., et al. (2005). Viral persistence in the myocardium is associated with progressive cardiac dysfunction. *Circulation* 112, 1965–1970. doi: 10.1161/circulationaha.105.548156

Lei, T., Lexun, L., Shuo, W., Zhiwei, G., Tianying, W., Ying, Q., et al. (2013). MiR-10a* up-regulates coxsackievirus B3 biosynthesis by targeting the 3D-coding sequence. *Nucleic Acids Res.* 41, 3760–3771. doi: 10.1093/nar/gkt058

Li, N., Hui, H., Bray, B., Gonzalez, G., Zeller, M., Anderson, K., et al. (2021). METTL3 regulates viral m6A RNA modification and host cell innate immune responses during SARS-CoV-2 infection. *Cell Rep.* 35:109091. doi: 10.1016/j.celrep.2021.109091

Li, X., Wang, Z., Ye, C., Zhao, B., Li, Z., and Yang, Y. (2018). RNA sequencing reveals the expression profiles of circRNA and indicates that circDDX17 acts as a tumor suppressor in colorectal cancer. *J. Exp. Clin. Cancer Res.* 37:325. doi: 10.1186/s13046-018-1006-x

Liang, Y., Han, H., Xiong, Q., Yang, C., Wang, L., Ma, J., et al. (2021). METTL3-mediated mA methylation regulates muscle stem cells and muscle regeneration by notch signaling pathway. *Stem Cells Int.* 2021:9955691. doi: 10.1155/2021/9955691

Lin, Q., Cai, J., and Wang, Q. (2020). The significance of circular RNA DDX17 in prostate cancer. *Biomed. Res. Int.* 2020, 1878431–1878416. doi: 10.1155/2020/1878431

Linder, P., and Jankowsky, E. (2011). From unwinding to clamping - the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.* 12, 505–516. doi: 10.1038/nrm3154

Liu, T., Tong, J., Shao, C., Qu, J., Wang, H., Shi, Y., et al. (2021). MicroRNA-324-3p plays a protective role against Coxsackievirus B3-induced viral myocarditis. *Virol. Sin.* 36, 1585–1599. doi: 10.1007/s12250-021-00441-4

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* 10, 93–95. doi: 10.1038/nchembio.1432

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Moy, R., Cole, B., Yasunaga, A., Gold, B., Shankarling, G., Varble, A., et al. (2014). Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense. *Cells* 158, 764–777. doi: 10.1016/j.cell.2014.06.023

Pauschinger, M., Chandrasekharan, K., Noutsias, M., Kühl, U., Schwimmbeck, L., and Schultheiss, H. (2004). Viral heart disease: molecular diagnosis, clinical prognosis, and treatment strategies. *Med. Microbiol. Immunol.* 193, 65–69. doi: 10.1007/s00430-003-0213-y

Peng, H., and Wen, Y. (2020). CircDDX17 acts as a competing endogenous RNA for miR-605 in breast cancer progression. *Eur. Rev. Med. Pharmacol. Sci.* 24, 6794–6801. doi: 10.26355/eurrev_202006_21668

Pinkert, S., Dieringer, B., Klopfleisch, R., Savvatis, K., Van Linthout, S., Pryshliak, M., et al. (2019). Early treatment of Coxsackievirus B3-infected animals with soluble Coxsackievirus-adenovirus receptor inhibits development of chronic Coxsackievirus B3 cardiomyopathy. *Circ. Heart Fail.* 12:e005250. doi: 10.1161/circheartfailure.119.005250

Ren, T., Liu, C., Hou, J., and Shan, F. (2020). CircDDX17 reduces 5-fluorouracil resistance and hinders tumorigenesis in colorectal cancer by regulating miR-31-5p/KANK1 axis. *Eur. Rev. Med. Pharmacol. Sci.* 24, 1743–1754. doi: 10.26355/eurrev_202002_20351

Roundtree, I., Evans, M., 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

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. (2011). A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? *Cells* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Shuo, W., Yan, W., Lexun, L., Xiaoning, S., Tianying, W., Xiaoyan, Z., et al. (2014). Protease 2A induces stress granule formation during coxsackievirus B3 and enterovirus 71 infections. *Virol. J.* 11:192. doi: 10.1186/s12985-014-0192-1

Terragni, J., Zhang, G., Sun, Z., Pradhan, S., Song, L., Crawford, G., et al. (2014). Notch signaling genes: myogenic DNA hypomethylation and 5-hydroxymethylcytosine. *Epigenetics* 9, 842–850. doi: 10.4161/epi.28597

Wang, F., Nazarali, A., and Ji, S. (2016). Circular RNAs as potential biomarkers for cancer diagnosis and therapy. *Am. J. Cancer Res.* 6, 1167–1176.

Wang, L., Xue, Y., Huo, R., Yan, Z., Xu, H., Li, H., et al. (2020). N6-methyladenosine methyltransferase METTL3 affects the phenotype of cerebral arteriovenous malformation via modulating notch signaling pathway. *J. Biomed. Sci.* 27:62. doi: 10.1186/s12929-020-00655-w

Winkler, R., Gillis, E., Lasman, L., Safra, M., Geula, S., Soyris, C., et al. (2019). mA modification controls the innate immune response to infection by targeting type I interferons. *Nat. Immunol.* 20, 173–182. doi: 10.1038/s41590-018-0275-z

Xia, T., Li, X., Wang, X., Zhu, Y., Zhang, H., Cheng, W., et al. (2021). N(6)-methyladenosine-binding protein YTHDF1 suppresses EBV replication and promotes EBV RNA decay. *EMBO Rep.* 22:e50128. doi: 10.15252/embr.202050128

Xiao, Y., Yang, Y., and Hu, D. (2021). Knockdown of METTL3 inhibits enterovirus 71-induced apoptosis of mouse Schwann cell through regulation of autophagy. *Pathog. Dis.* 79:6. doi: 10.1093/femspd/ftab036

Xu, J., Cai, Y., Ma, Z., Jiang, B., Liu, W., Cheng, J., et al. (2021). The RNA helicase DDX5 promotes viral infection via regulating N6-methyladenosine levels on the DHX58 and NFκB transcripts to dampen antiviral innate immunity. *PLoS Pathog.* 17:e1009530. doi: 10.1371/journal.ppat.1009530

Yang, Q., Li, Y., Wang, Y., Qiao, X., Liu, T., Wang, H., et al. (2021). The circRNA circSIAE inhibits replication of Coxsackie virus B3 by targeting miR-331-3p and thousand and one amino-acid kinase 2. *Front. Cell. Infect. Microbiol.* 11:779919. doi: 10.3389/fcimb.2021.779919

Zaccara, S., Ries, R., and Jaffrey, S. (2019). Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* 20, 608–624. doi: 10.1038/s41580-019-0168-5

Zhao, B., Roundtree, I., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* 18, 31–42. doi: 10.1038/nrm.2016.132

Zhao, C., Yan, Z., Wen, J., Fu, D., Xu, P., Wang, L., et al. (2021). CircEAF2 counteracts Epstein-Barr virus-positive diffuse large B-cell lymphoma progression via miR-BART19-3p/APC/β-catenin axis. *Mol. Cancer* 20:153. doi: 10.1186/s12943-021-01458-9

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fmicb.2022.1030545

# Surveillance of avian influenza viruses in live bird markets of Shandong province from 2013 to 2019

Ti Liu[1†], Yousong Peng[2†], Julong Wu[1], Shangwen Lu[2], Yujie He[1], Xiyan Li[3], Lin Sun[1], Shaoxia Song[1], Shengyang Zhang[1], Zhong Li[1], Xianjun Wang[1], Shu Zhang[1], Mi Liu[4]* and Zengqiang Kou[1]*

[1]Shandong Provincial Key Laboratory of Infectious Disease Control and Prevention, Shandong Center for Disease Control and Prevention, Jinan, China, [2]Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha, China, [3]Chinese National Influenza Center, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China, [4]Jiangsu Institute of Clinical Immunology, The First Affiliated Hospital of Soochow University, Suzhou, China

Avian influenza viruses (AIVs) seriously affect the poultry industry and pose a great threat to humans. Timely surveillance of AIVs is the basis for preparedness of the virus. This study reported the long-term surveillance of AIVs in the live bird market (LBM) of 16 cities in Shandong province from 2013 to 2019. A total of 29,895 samples were obtained and the overall positive rate of AIVs was 9.7%. The H9 was found to be the most predominant subtype in most of the time and contributed most to the monthly positve rate of AIVs as supported by the univariate and multivariate analysis, while H5 and H7 only circulated in some short periods. Then, the whole-genome sequences of 62 representative H9N2 viruses including one human isolate from a 7-year-old boy in were determined and they were genetically similar to each other with the median pairwise sequence identities ranging from 0.96 to 0.98 for all segments. The newly sequenced viruses were most similar to viruses isolated in chickens in mainland China, especially the provinces in Eastern China. Phylogenetic analysis showed that these newly sequenced H9N2 viruses belonged to the same clade for all segments except PB1. Nearly all of these viruses belonged to the G57 genotype which has dominated in China since 2010. Finally, several molecular markers associated with human adaptation, mammalian virulence, and drug resistance were identified in the newly sequenced H9N2 viruses. Overall, the study deepens our understanding of the epidemic and evolution of AIVs and provides a basis for effective control of AIVs in China.

KEYWORDS

**avian influenza viruses, surveillance, H9N2 AIV, epidemic, genotypes**

OPEN ACCESS

EDITED BY
Zhan Zhou,
Zhejiang University,
China

REVIEWED BY
Mohammad Mahmudul Hassan,
The University of Queensland, Australia
Hongxing Shen,
Jiangsu University, China

*CORRESPONDENCE
Zengqiang Kou
jack-kou@163.com
Mi Liu
liumi@suda.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

RECEIVED 29 August 2022
ACCEPTED 11 October 2022
PUBLISHED 03 November 2022

CITATION
Liu T, Peng Y, Wu J, Lu S, He Y, Li X, Sun L, Song S, Zhang S, Li Z, Wang X, Zhang S, Liu M and Kou Z (2022) Surveillance of avian influenza viruses in live bird markets of Shandong province from 2013 to 2019. *Front. Microbiol.* 13:1030545.
doi: 10.3389/fmicb.2022.1030545

## Introduction

The influenza A virus belongs to the *Orthomyxoviridae* family and contains a negative-sense RNA genome with eight segments. The influenza A virus is classified into different subtypes based on the surface proteins of haemagglutinin (HA) and neuraminidase (NA), such as H7N9, H5N1, or H9N2. The natural hosts of the influenza A virus are wild birds including both the wild waterfowl and sea birds (Webster et al., 1992). The avian influenza viruses (AIVs) are influenza A viruses that mainly infect the avians including both wild birds and poultry. The AIVs have caused numerous epidemics in poultry globally and have seriously affected the poultry industry (Su et al., 2015). Besides, they can occasionally cause human infections (Mostafa et al., 2018). Lots of subtypes of AIVs have been reported to infect humans in recent years, such as H5N1, H5N6, H7N9, H9N2, H10N8, H5N8, H3N8, and so on (Mostafa et al., 2018; Li et al., 2019b). How to effectively control AIVs is a great challenge for humans.

China is among the countries with the most diverse AIVs as the country has a large number of wild bird species and maintains the largest number of poultry in the world (Liu S. et al., 2020). Multiple subtypes of AIVs have circulated extensively in China in the last 20 years. Among them, the subtypes of H5, H7, and H9 are most predominant in China (Su et al., 2015; Liu S. et al., 2020). The subtype H5 has circulated in China for more than 20 years since several large-scale outbreaks occurred in 2001 which were caused by the highly pathogenic avian influenza (HPAI) H5N1 virus (Peng et al., 2017). The HPAI H5N1 virus has ever been considered to be most likely to cause global pandemics before 2009 when the H1N1 pandemic happened. It was almost the exclusive subtype among H5 subtypes which circulated in China before 2012 (Liu S. et al., 2020). Then, the subtypes of H5N2, H5N6, and H5N8 emerged and replaced the HPAI H5N1 virus in China. The subtype H7 has been widely circulating in China since 2013 when the H7N9 virus caused human infections (Jiang et al., 2019). Until now, the H7N9 virus has caused more than 1,500 confirmed human infections and more than 500 human deaths in China (Quan et al., 2018). The virus has caused multiple outbreaks in chickens since it evolved into the HPAI virus in 2016. Fortunately, the virus is now rarely detected in China because of the simultaneous immunization of the H5 + H7 vaccine among poultry since 2017 (Li and Chen, 2021).

Compared to subtypes of H5 and H7, the H9 subtype was the first AIV subtype that caused widespread infections in poultry in China (Peacock et al., 2019; Liu S. et al., 2020). The H9N2 virus has been circulating in China since the 1990s. Several large-scale surveillance studies have shown that the H9N2 virus was the most prevalent subtype in poultry in China (Bi et al., 2020). The virus also caused sporadic human infections, most of which happened in poultry workers. The H9N2 virus in China could be classified into three large clades, i.e., BJ/94, G1, and F/98 based on epidemiological and phylogenetic analysis. Li et al. further classified the H9N2 virus in China into at least 117 genotypes by evolutionary analysis (Li et al., 2017). Among them, the G57 genotype has become dominant in China since 2010 (Peacock et al., 2019). Due to the large diversity and high prevalence in birds, the H9N2 virus plays an important role in the evolution of AIVs by providing internal genes to other AIVs, which lead to novel AIVs (Peacock et al., 2019). For example, the H7N9 virus which has caused human infections since 2013 was reported to obtain all its internal genes from H9N2 viruses by re-assortment (Wu et al., 2013).

Shandong is a big agricultural province of China and has a developed poultry industry which poses a high risk of AIV outbreaks. Timely surveillance of AIV is the basis for better preparedness of the virus. However, there was little data about the epidemiology and evolution of AIVs in the province in recent years. This study reported the long-term surveillance of AIVs in the live bird market (LBM) of 16 cities in Shandong province from 2013 to 2019. The H9 was found to be the most predominant subtype during the period. Thus, the whole-genome sequences of 62 representative H9N2 viruses including 61 environmental isolates and one human isolates were determined by the next-generation-sequencing method. The evolution of these viruses and the molecular markers they contained were further analyzed. The study deepens our understanding of the epidemic and evolution of AIVs and provides a basis for effective control of AIVs in China.

## Materials and methods

### Virus sampling and isolation

A total of 29,895 samples were obtained from the environments including the surface wipe of poultry cages, chopping boards, poultry drinking water and feces, of LBMs in 16 cities of Shandong province, China, from 2013 to 2019. The samples were placed in the 3 ml of Viral Transport Medium (VTM), and then were centrifuged at 3000 g for 10 min. The supernatants were used to extract RNA with the Qiagen RNeasy Mini Kit (Lot. 74,104) according to the manufacturer's instructions. The real-time RT-PCR was used to detect AIVs. If the sample was positive for AIVs, the sample was further subtyped for H5, H7 and H9. The positive samples of H9 subtype were inoculated into the allantoic cavity of 9-day-old specific pathogen-free embryonated chicken eggs and incubated for 72 h at 37°C and chilled at 4°C overnight. The allantoic fluids were harvested and the influenza A(H9N2) virus strains were identified with a combination of hemagglutination assay with horse erythrocytes and real-time RT-PCR of H9N2 detection.

### Human H9N2 case finding and isolation

On April 28th, 2020, a 7-year-old boy living in Weihai, a city in Shandong province, was taken to Weihai Municipal Hospital for influenza-like illness. A nasopharyngeal sample obtained was tested positive for influenza A and H9N2 at the Weihai Center for Disease Control and Prevention. The sample was then sent to the Shandong Provincial Center for Disease Control and Prevention,and the virus

(A/human/shandong/01/2020) was successfully isolated in embryonated chicken eggs and confirmed as H9N2.

## Genome sequencing

A total of 62 H9N2 strains including 61 environmental strains isolated in the surveillance efforts and one human isolate mentioned above were sequenced by the next-generation-sequencing method (Supplementary Table S1). The total viral RNA of these strains was extracted with the Qiagen RNeasy Mini Kit (Lot. 74,104). The RNA was subjected to reverse transcription and amplification using the SuperScript™ III One-Step RT-PCR System with Platinum™ Taq High Fidelity DNA Polymerase (cat#: 12574035, Invitrogen). The DNA library was prepared using Nextera XT DNA Preparation Kits (cat#FC-131-1,096, Illumina). Whole-genome sequencing was then performed on MiSeq high-throughput sequencing platform (Illumina, Inc., San Diego, CA, United States), and the data were analyzed using CLC Genomics Workbench software.

## Phylogenetic analysis

Except for newly sequenced H9N2 viruses, we also collected other H9N2 viruses from the public database for phylogenetic analysis. The nucleotide sequences of H9N2 viruses were downloaded from the database of Influenza Virus Resource on October 25th, 2021 (Bao et al., 2008), and were clustered using CD-HIT to create the reference sequence database (Li and Godzik, 2006). The representative viruses together with the newly sequenced viruses were used to build phylogenetic trees. Phylogenetic trees for all eight segments of H9N2 viruses were generated by the maximum likelihood method using MEGA X (Kumar et al., 2018). The neighbor virus of the newly sequenced virus on each segment was identified by querying against the reference sequence database using BLASTN (version 2.13.0) (Altschul et al., 1997). The virus strain with the highest bit score was selected as the neighbor virus of the query sequence.

## Genotype determination

The genotypes of newly sequenced H9N2 viruses were determined based on the phylogenetic analysis according to previous studies (Pu et al., 2015; Li et al., 2017; Jin et al., 2020).

## Identification of molecular markers in H9N2 viruses

The human-adaptation, mammalian virulence and drug-resistance associated molecular markers in newly sequenced H9N2 viruses were identified with FluPhenotype on July 12th, 2022 (Lu et al., 2020).

## Statistical analysis

The univariate and multivariate analysis of the monthly positive rate of influenza viruses by subtype was conducted using the "lm" function in R (version 3.6.1).

# Results

## Surveillance of AIVs in LBMs of Shandong province

A total of 29,895 samples from the environments such as the surface wipe of poultry cages, poultry drinking water and feces in LBMs of 16 cities in Shandong province, China, were obtained from January 2013 to April 2019 (Figure 1A). Among them, 2,903 samples were positive for AIVs and the overall positive rate was 9.7%. When analyzed by city, the number of samples surveyed ranged from 150 to 4,977 in 16 cities. The positive rate ranged from 1 to 22% in these cities, with Binzhou, Heze and Weifang having the highest positive rates (Figure 1A).

The overall monthly positive rate of AIVs ranged from 0 to 0.485 with a median of 0.052 from January 2013 to April 2019 (Figure 1B and Supplementary Table S2). When analyzed by HA subtype, the monthly positive rates of H5, H7 and H9 were calculated from 2013 to 2019. Attention to note, the positive rate during the period of 2013–2015 may be underestimated as a large portion of samples was un-subtyped during the period. Both the univariate and multivariate analysis showed that the H9 subtype contributed most to the monthly positive rate (Table 1), while other HA subtypes or subtype combinations contributed little. The year and month had minor positive and negative effects, respectively, on the monthly positive rate in the univariate analysis, although both had no statistical significance. Interestingly, they became statistically significant in the multivariate analysis, suggesting complex interactions between subtype, year and month.

As shown in Figure 1B, H9 was the main subtype in most of the time (Figure 1B), which was consistent with the univariate and multivariate analysis. It persisted circulating throughout the year from 2013 to 2019. In general, the H9 peaked in the winters and maintained a low level of circulation in the summers. H5 mainly circulated during the season of 2015–2016. It caused sporadic infections in some months. H7 circulated least compared to H5 and H9. It only dominated in the April of 2017 when 20 cases of human infections of the H7N9 virus were reported in Shandong province in the same year (Zhang et al., 2020), and maintained a very low level of activity in most of the time.

Besides the positive rates of individual HA subtypes, we also surveyed the co-occurrence of H5, H7, and H9 during the period. As expected, most co-occurrences happened between H9 and H5 or H7, and the co-occurrences were observed when both HA subtypes had high positive rates such as the co-occurrence of H9 and H5 observed in early 2016. Only a few co-occurrences of H5 and H7 were observed from 2013 to 2019 (Figure 1B).

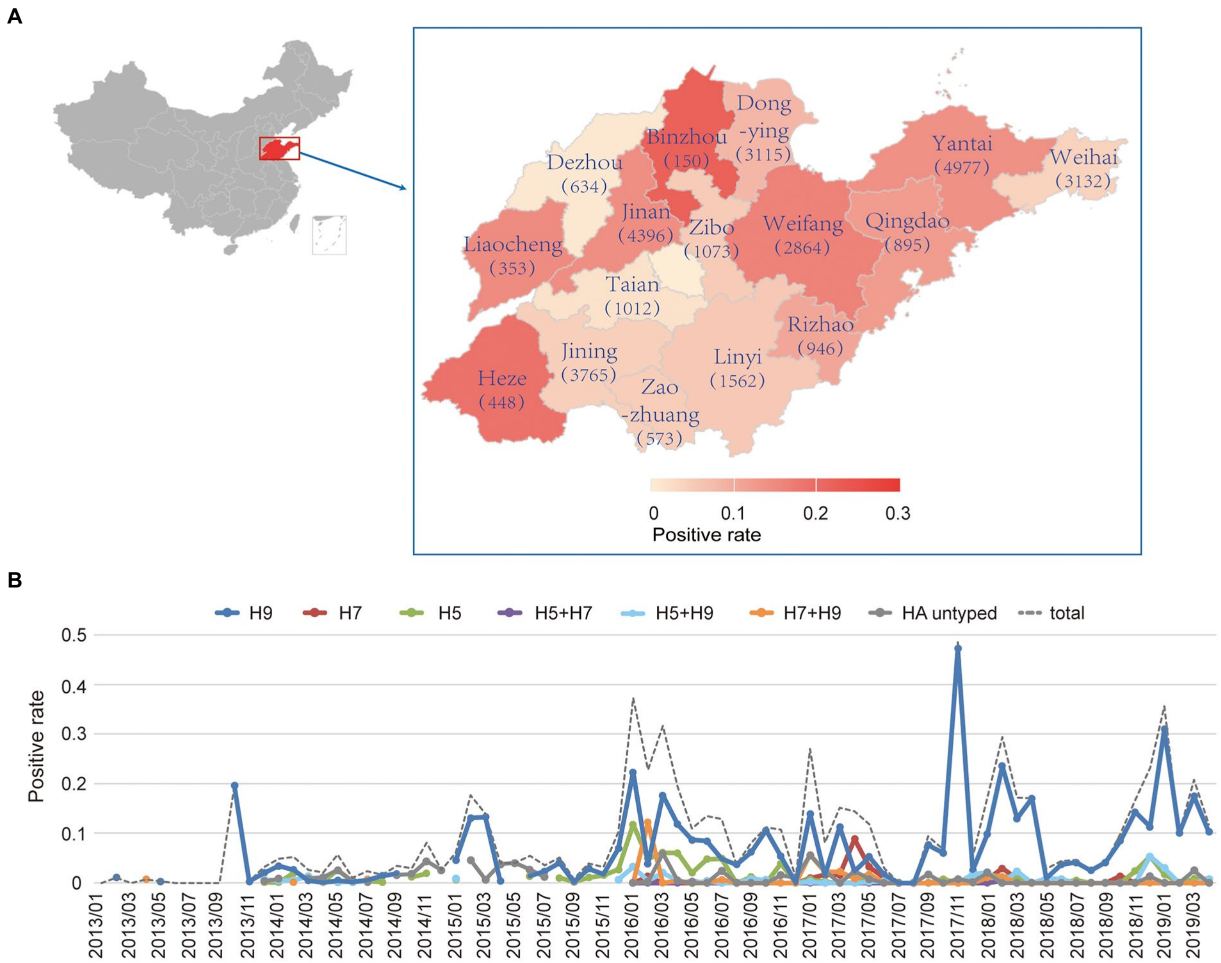


FIGURE 1
Surveillance of AIVs in LBMs of Shandong province from 2013 to 2019. **(A)** The number of samples and positive rates of AIVs by city in Shandong province. The cities were colored by the positive rate. The numbers in parentheses referred to the number of samples surveyed in the city. **(B)** The monthly positive rate of different HA subtypes or subtype combinations of influenza A viruses.

TABLE 1 **Univariate and multivariate analysis of the monthly positive rate by subtype.**

| Variable | Univariate analysis | | | Multivariate analysis | | |
|---|---|---|---|---|---|---|
| | Coefficient | Standard Error | *p*-value | Coefficient | Standard Error | *p*-value |
| *Subtype* | | | | | | |
| H9 | 0.060 | 0.007 | 2.39E-16 | 0.061 | 0.007 | <2.2E-16 |
| H7 | −0.005 | 0.008 | 0.509 | | | |
| H5 | 0.003 | 0.007 | 0.651 | | | |
| H5&H7 | −0.011 | 0.008 | 0.169 | | | |
| H5&H9 | −0.005 | 0.007 | 0.504 | | | |
| H7&H9 | −0.004 | 0.008 | 0.576 | | | |
| Year | 0.0006 | 0.002 | 0.714 | 0.004 | 0.001 | 0.0098 |
| Month | −0.001 | 0.0007 | 0.070 | −0.001 | 0.0006 | 0.042 |

## Sequencing and phylogenetic analysis of representative H9N2 viruses

Since the H9 was the most dominant subtype in our surveillance, the whole genome sequences of 62 representative H9N2 strains including 61 environment strains and one human isolate were obtained using the next-generation-sequencing method for better understanding the genetic evolution of H9N2 viruses in Shandong province (Supplementary Table S1). As shown in Figure 2, for all segments, the median pairwise sequence

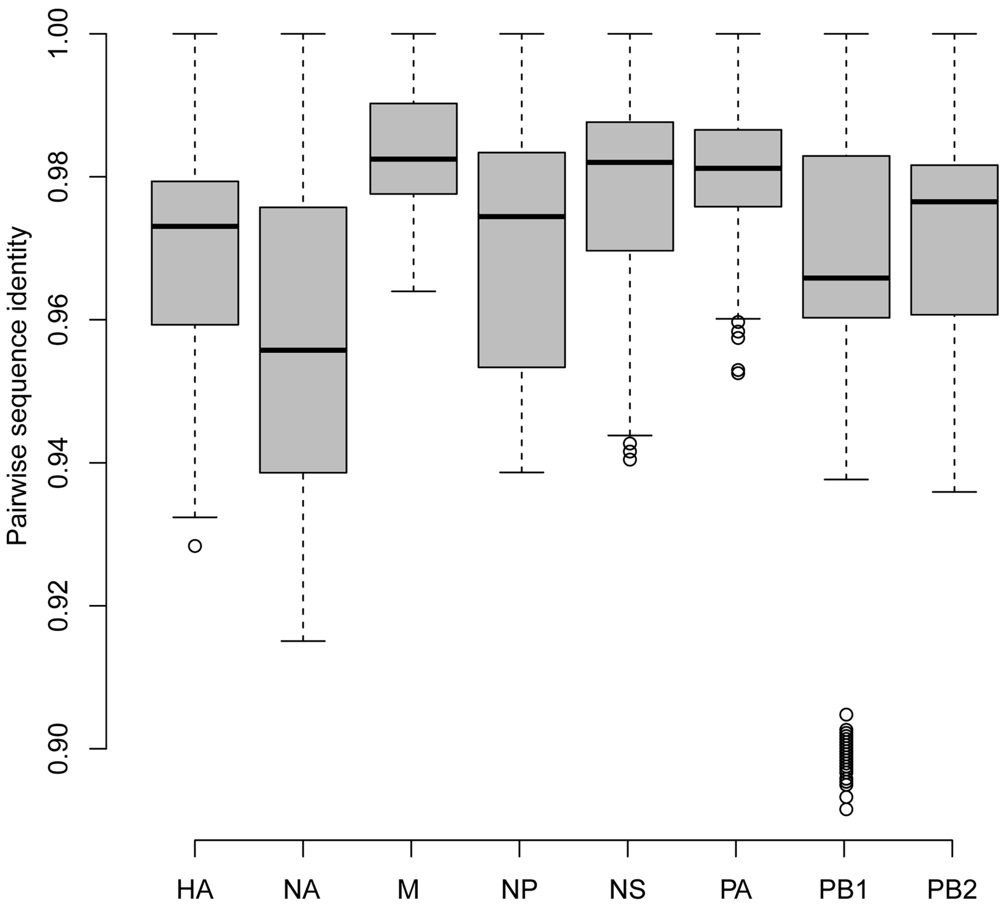


FIGURE 2
Pairwise sequence identities of each segment between newly sequenced H9N2 viruses.

identities ranged from 0.96 to 0.98. The gene M had the highest pairwise sequence identities, while the NA gene had the lowest pairwise sequence identities. The HA gene had pairwise sequence identities ranging from 0.93 to 1, with a median of 0.97.

These viruses were further genetically characterized by the phylogenetic analysis. For each segment of H9N2 viruses, several reference viruses of influenza H9N2 viruses were selected from the Influenza Virus Resource database and were used in phylogenetic analysis with the newly sequenced H9N2 viruses (see Materials and Methods). As shown in Figure 3, for all segments except PB1, all newly sequenced H9N2 strains (colored in red) belonged to the same clade (marked with a dashed box). For PB1, all viruses belong to clade 5 except the viral isolate A/Environment/shandong-dongying/08/2015 which belonged to clade 2 and was clustered with sequences isolated from wild birds (Supplementary Figure S1).

We then investigated the possible source of the newly sequenced H9N2 viruses since all of them except the human isolate were isolated from the environment. For each segment of each virus, the neighbor of the virus (defined as the most similar virus) was obtained (see Materials and Methods), and the composition of the host, isolation location and year of the neighbors were analyzed (Figure 4). In terms of host, for all segments, most neighbors were isolated from the poultry including chicken and ducks, especially the chicken. In terms of isolation location, nearly all neighbors were isolated from mainland China, especially the Shandong province and its neighboring provinces in Eastern China such as Anhui, Jiangsu, and Zhejiang provinces. In terms of isolation time, most neighbors were isolated in the same year with, or 1 year before or after the year when the virus was isolated.

The neighbors of the human isolate A/human/shandong/01/2020 were analyzed individually. In terms of host, the neighbors of the human isolate were isolated from chicken for all segments; in terms of isolation location, for five segments (HA, NA, MP, PA and PB2), the neighbors were isolated in Anhui province, while for NP, NS and PB1, the neighbors were isolated in Fujian, Jiangxi and Shanxi province, respectively; in terms of isolation time, the neighbors of the human isolate were isolated in 2018 for all segments except NA of which the neighbor was isolated in 2019.

## Genotyping of newly sequenced H9N2 viruses

The genotypes of newly sequenced H9N2 viruses were determined based on the phylogenetic analysis according to Li's

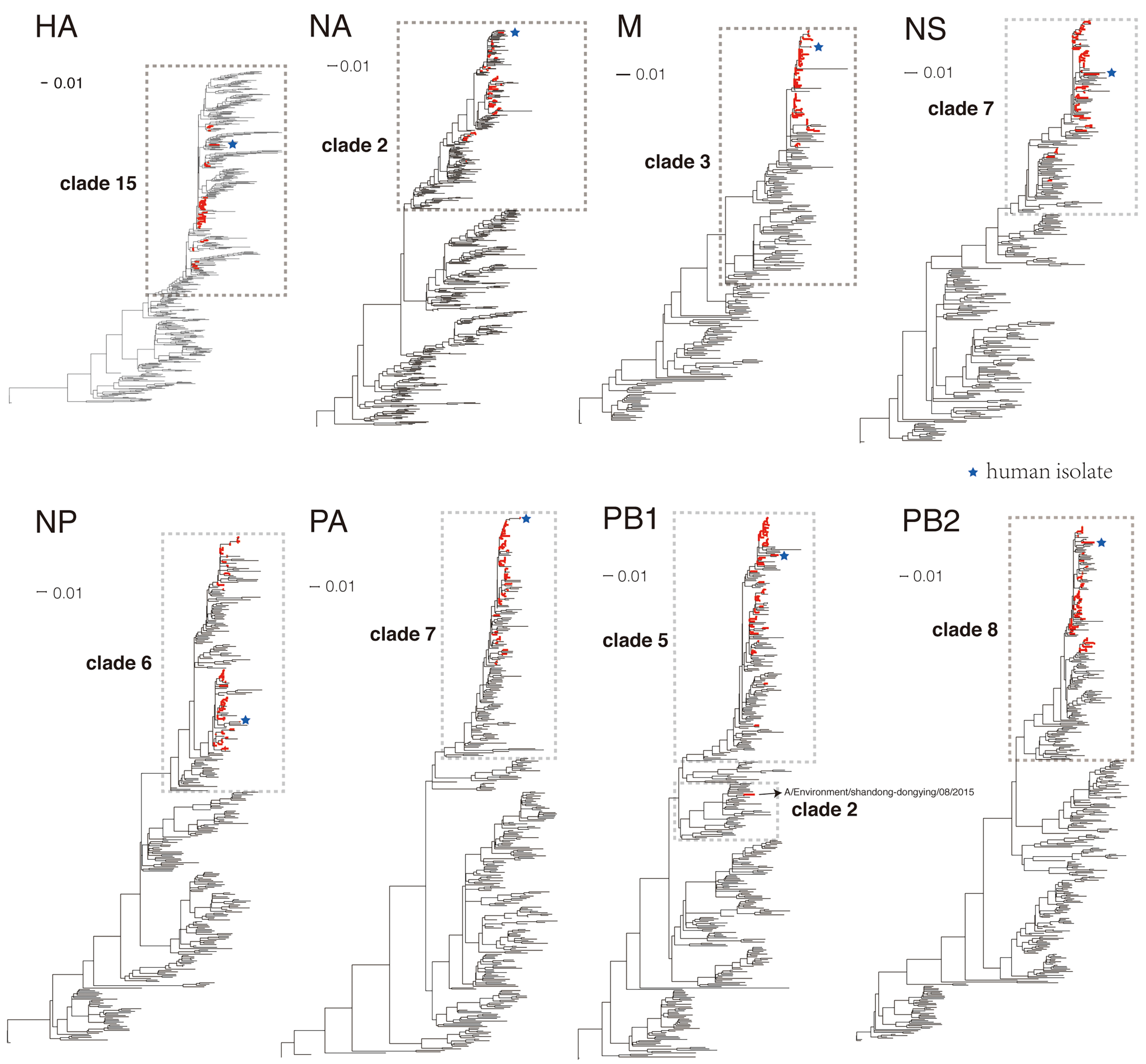


FIGURE 3
The phylogeny of H9N2 viruses on each segment. The newly sequenced H9N2 viruses were colored in red in the trees. The human isolate was marked with blue stars. The details of these trees were shown in Supplementary Figure S1.

study (Supplementary Table S3) (Materials and Methods). All viruses except A/Environment/shandong-dongying/08/2015 belonged to the G57 genotype which was first found in Eastern China in 2007 and has been the predominant genotype in China since 2010 (Li et al., 2017). The A/Environment/shandong-dongying/08/2015 belonged to G69 genotype which has a different PB1 clade compared to G57 and was rarely reported in China.

## Molecular markers of H9N2 viruses

Finally, we investigated the molecular markers associated with human adaptation, mammalian virulence, and drug resistance in the newly sequenced H9N2 viruses using FluPhenotype (Lu et al., 2020) (Materials and Methods). In terms of human adaptation, 11 molecular markers that were located in 7 proteins were observed, and 7 of them happened in more than 50 viruses (Table 2). For example, Threonine on 180 of the HA protein that was reported to be associated with an increase in binding to the human-like receptor was observed in 51 of 62 newly sequenced H9N2 viruses. The human isolate A/human/shandong/01/2020 had 6 human-adaptation associated molecular markers including HA-180 T, HA2-46E, PA-356R, PB1-368 V, PB1-F2-47S and PB2-66 M.

In terms of mammalian virulence, 30 molecular markers that were located in 6 proteins were observed and 18 of them happened in more than 50 viruses (Table 3). For example, three markers in the MP1 protein including 30D, 43 M and 215A which were reported to increase virulence in mammals were observed in all

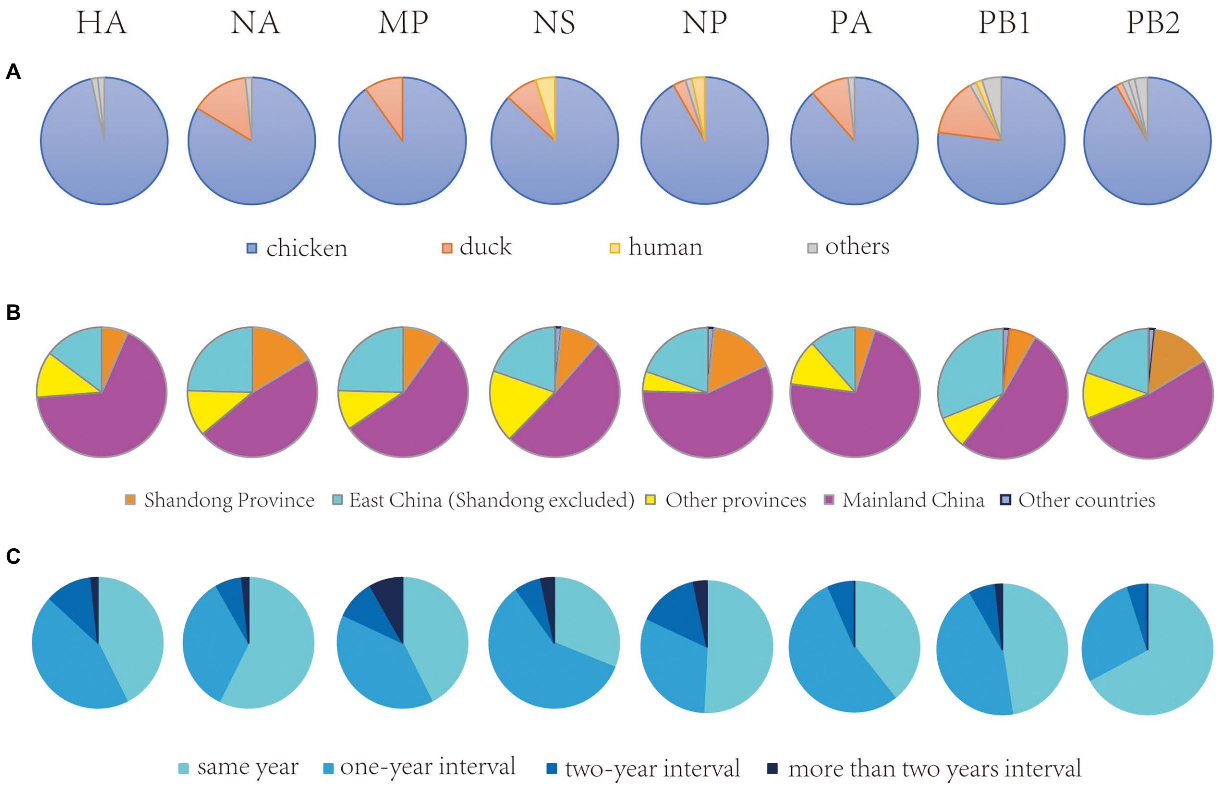


FIGURE 4
The host, isolation location and year composition of the neighbors of newly sequenced H9N2 viruses isolated in the environment.

TABLE 2 The human-adaptation-related molecular markers identified in the newly sequenced H9N2 virus strains.

| Protein | Mutation | Ratio | Mutation effect |
|---|---|---|---|
| HA | **180 T** | 51/62 | Increase in binding to the human-like receptor |
| | **HA2-46E** | 62/62 | Enhanced binding to both avian-and human-type receptor |
| M2 | 19Y | 2/62 | The human-adaptation associated residues |
| NP | 109 V | 1/62 | The human-adaptation associated residues |
| PA | 57Q | 1/62 | The human-adaptation associated residue |
| | 100A | 2/62 | The human-adaptation associated residue |
| | **356R** | 62/62 | The human-adaptation associated residue |
| PB1 | **368 V** | 55/62 | The human-adaptation associated residue |
| PB1-F2 | **47S** | 62/62 | The human-adaptation associated residue |
| PB2 | **66 M** | 61/62 | The human-adaptation associated residue |
| | **89 V** | 62/62 | The human-adaptation associated residue |

Those which were found in more than 50 viruses were highlighted in bold.

62 newly sequenced H9N2 viruses. The human isolate A/human/shandong/01/2020 had 6 molecular markers associated with mammalian virulence including MP1-30D, MP1-43 M, MP1-215A, NS1-42S, PA-224S and PB2-431 M.

In terms of drug resistance, 4 molecular markers which were located in NA and MP2 proteins were observed (Supplementary Table S4). The molecular marker NA-151D which was reported to be associated with resistance to oseltamivir and zanamivir, and molecular markers of MP2-21G and MP2-31 N which were reported to be associated with resistance to amantadine were observed in more than 60 newly sequenced H9N2 strains.

## Discussion

The LBM has been reported to play an important role in the spreading of AIVs because numerous poultry are transported in and out of LBMs (Cardona et al., 2009; Yu et al., 2014; Li et al., 2018).

TABLE 3 The mammalian-virulence-related molecular markers identified in the newly sequenced H9N2 strains.

| Gene | Mutation | Ratio | Mutation effect |
|---|---|---|---|
| MP1 | **30D** | 62/62 | Increased virulence in mice |
| | **43 M** | 62/62 | Contribute to the pathogenicity of HPAI H5N1 viruses in both avian and mammalian hosts |
| | **215A** | 62/62 | Increased virulence in mice |
| NS1 | **42S** | 61/62 | Increased virulence in mice |
| | 123 V | 5/62 | Selective advantage in replication and/or transmission of pH1N1 in humans |
| | 127 N | 40/62 | Associated with high-virulence in mammals |
| PA | **37S** | 62/62 | Decreased viral transcription and replication by diminishing virus RNA synsthesis activity |
| | **37&61ST** | 62/62 | Decreased viral transcription and replication by diminishing virus RNA synthesis activity |
| | **63I** | 62/62 | Decreased viral transcription and replication by diminishing virus RNA synthesis activity |
| | 70&224VS | 17/62 | Reducing the virus LD50 in mice by almost 1,000-fold |
| | **190S** | 62/62 | Reduced the virulence of H7N3 virus |
| | **224S** | 62/62 | Enhanced polymerase activity and virulence of pH1N1 in mice |
| | **400P** | 62/62 | Reduced the virulence of H7N3 virus |
| | 409S | 2/62 | Increased virus replication ability in mammalian systems |
| | **550l** | 62/62 | Increased polymerase activity and high virulence. |
| PB1 | **13P** | 62/62 | Enhanced polymerase activities of 270% |
| | **317I** | 56/62 | I at 317 in PB1 correlated with high pathogenicity. |
| PB1-F2 | 51&56&87TVE | 1/62 | Increased viral polymerase activity and expression levels of viral RNA |
| PB2 | 195 N | 5/62 | The PB2-D195 N substitution increased polymerase activity by about 3.5-fold |
| | 283&526MR | 3/62 | Enhanced virulence of H5N8 influenza viruses in mice |
| | **292 V** | 54/62 | Higher viral polymerase activity and stronger attenuation of host IFN-Î2 response |
| | **309D** | 62/62 | Increased polymerase activity |
| | **431 M** | 62/62 | Impacts the viral replication and virulence in mice by altering the viral polymerase activity |
| | **504 V** | 62/62 | Mutational analyses demonstrated that an isoleucine-to-valine change at position 504 in PB2 was the most critical and strongly enhanced the activity of the reconstituted polymerase complex. |
| | 526R | 3/62 | Increased polymerase activity/enhanced pathogenicity in mice |
| | 535l | 12/62 | Increased polymerase activity |
| | 598I | 10/62 | Increased virus replication and virulence in mice |
| | **661A** | 60/62 | Increased polymerase activity at low temperature |
| | 702R | 11/62 | Increased polymerase activity / enhanced pathogenicity in mice |
| | 89&309&339&477&495&627&676 VDKGVET | 1/62 | Increased virulence in mice. |

Those which were found in more than 50 viruses were highlighted in bold.

Studies have shown that closing the LBM had a large impact on the human infection of AIVs such as H7N9 viruses (Yu et al., 2014; Li et al., 2018). Besides, the LBM was also considered to be a vessel for mixing AIVs (Cardona et al., 2009). Lots of novel viruses can be generated in the LBM by re-assortment such as the H7N9 virus which has caused human infections since 2013 (Wu et al., 2013). In this study, we surveyed the AIV in LBMs of Shandong province from 2013 to 2019 and found persistent circulations of AIVs in the province. The H9 was found to be the main subtype in most of the time, while the H5 and H7 only dominated in some short periods. These were consistent with previous studies which showed the predominant role of H9 in China (Peacock et al., 2019; Liu S. et al., 2020). Our analysis also found significant co-occurrence of H5, H7 and H9, suggesting the high risk of re-assortment of AIVs in the LBM. Therefore, continual surveillance of AIVs in the LBM is required for the timely identification of novel flu viruses.

Vaccination is the best way for controlling the avian influenza virus. Large scale vaccination of poultry has been conducted in China for prevention and control of the avian influenza viruses including H5, H9 and H7 (Liu S. et al., 2020). On one hand, vaccination can greatly stop the spreading of avian influenza virus. For example, massive vaccination of chickens with an H5/H7 bivalent avian influenza vaccine since September 2017 has successfully controlled H7N9 avian influenza infections in poultry (Li and Chen, 2021). Our surveillance also showed that few H5 or H7 epidemics were found in Shandong provinces since 2018. One the other hand, vaccination pressure can drive the antigenic evolution of avian influenza viruses such as H5 virus (Peng et al., 2017). More surveillance of the antigenic variation of both H5 and H7 viruses are needed to capture the antigenic variant in time.

A total of 62 H9N2 viruses were sequenced in the study. They were found to be highly similar to each other in all segments. All of these viruses except one isolate belonged to the G57 genotype which has been the dominant genotype of H9N2 viruses circulating in China since 2010. This is consistent with Li's study which showed that all nine strains of H9N2 viruses isolated in chicken flocks in Shandong province in 2018 belonged to the G57 genotype (Li et al., 2019a). This suggested the great advantage of G57 compared to other genotypes in China. Most newly sequenced H9N2 viruses were similar to those isolated in chickens in mainland China, suggesting the prevalence of H9N2 viruses in chickens.

Molecular markers play important roles in surveillance of emerging influenza viruses, such as monitoring the antigenic variation, drug resistance and host adaptation of the virus (Chen et al., 2006; Liu W. J. et al., 2020). Lots of molecular markers of antigen, host, pathogenicity and drug-resistance have been identified for influenza viruses (Lu et al., 2020; Peng et al., 2020). In the study, all newly sequenced H9N2 viruses harbored several molecular markers associated with human adaptation, mammalian virulence and drug resistance. Although most molecular markers have been experimentally-validated, the role of them in the newly sequenced H9N2 viruses may be changed due to the epistasis (Lyons and Lauring, 2018). Further experiments are needed to validate their role in the newly sequenced H9N2 viruses. Interestingly, the human H9N2 isolate did not have more molecular markers than other viruses, suggesting that the isolate may infect humans accidentally. Nevertheless, more strict protective measures against the H9N2 infection are needed for high-risk people such as poultry worker.

Overall, this study systematically surveyed the AIVs in the LBM of Shandong province from 2013 to 2019 and further revealed the diversity and evolution of H9N2 viruses in the province. It deepens our understanding of the epidemic and evolution of AIVs, and would greatly facilitate the prevention and control of AIVs in China.

## Data availability statement

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

## Ethics statement

The studies involving human participants were reviewed and approved by IRB for Preventive Medicine of Shandong Center for Disease Control and Prevention (reference 2021–24).

## Author contributions

Conceptualization, ZK, ML, TL, and YP: Methodology, TL, YP: Formal Analysis, YP, ML: Investigation, XL, LS and SS: Resources, JL, SL, YH, SZ, ZL, XW and SZ: Writing – Original Draft, TL, YP, ML: Writing – Review and Editing, ZK, TL, YP and ML: Funding Acquisition, TL and YP. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

This work was supported by Shandong provincial Natural Science Foundation (ZR2021MH372) and the National Natural Science Foundation of China (32170651).

## 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.2022.1030545/full#supplementary-material

## References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Bao, Y. M., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., et al. (2008). The influenza virus resource at the national center for biotechnology information. *J. Virol.* 82, 596–601. doi: 10.1128/JVI.02005-07

Bi, Y., Li, J., Li, S., Fu, G., Jin, T., Zhang, C., et al. (2020). Dominant subtype switch in avian influenza viruses during 2016-2019 in China. *Nat. Commun.* 11:5909. doi: 10.1038/s41467-020-19671-3

Cardona, C., Yee, K., and Carpenter, T. (2009). Are live bird markets reservoirs of avian influenza? *Poult. Sci.* 88, 856–859. doi: 10.3382/ps.2008-00338

Chen, G.-W., Chang, S.-C., Mok, C., Lo, Y.-L., Kung, Y.-N., Huang, J.-H., et al. (2006). Genomic signatures of human versus avian influenza a viruses. *Emerg. Infect. Dis.* 12, 1353–1360. doi: 10.3201/eid1209.060276

Jiang, W., Hou, G., Li, J., Peng, C., Wang, S., Liu, S., et al. (2019). Prevalence of H7N9 subtype avian influenza viruses in poultry in China, 2013-2018. *Transbound. Emerg. Dis.* 66, 1758–1761. doi: 10.1111/tbed.13183

Jin, X., Zha, Y., Hu, J., Li, X., Chen, J., Xie, S., et al. (2020). New molecular evolutionary characteristics of H9N2 avian influenza virus in Guangdong Province, China. *J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 77:104064. doi: 10.1016/j.meegid.2019.104064

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.

Li, C., and Chen, H. (2021). H7N9 Influenza Virus in China. *Cold Spring Harb. Perspect. Med.* 11:a038349. doi: 10.1101/cshperspect.a038349

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, Y. T., Linster, M., Mendenhall, I. H., Su, Y. C. F., and Smith, G. J. D. (2019b). Avian influenza viruses in humans: lessons from past outbreaks. *Br. Med. Bull.* 132, 81–95. doi: 10.1093/bmb/ldz036

Li, Y. T., Liu, M., Sun, Q., Zhang, H., Jiang, S., et al. (2019a). Genotypic evolution and epidemiological characteristics of H9N2 influenza virus in Shandong Province, China. *Poult. Sci.* 98, 3488–3495. doi: 10.3382/ps/pez151

Li, C., Wang, S., Bing, G., Carter, R. A., Wang, Z., Wang, J., et al. (2017). Genetic evolution of influenza H9N2 viruses isolated from various hosts in China from 1994 to 2013. *Emerg. Microb. Infect.* 6:e106. doi: 10.1038/emi.2017.94

Li, Y., Wang, Y., Shen, C., Huang, J., Kang, J., and Huang, B. (2018). Closure of live bird markets leads to the spread of H7N9 influenza in China. *PloS One* 13:e0208884. doi: 10.1371/journal.pone.0208884

Liu, W. J., Li, J., Zou, R., Pan, J., Jin, T., Li, L., et al. (2020). Dynamic PB2-E627K substitution of influenza H7N9 virus indicates the in vivo genetic tuning and rapid host adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 117, 23807–23814.

Liu, S., Zhuang, Q., Wang, S., Jiang, W., Jin, J., Peng, C., et al. (2020). Control of avian influenza in China: strategies and lessons. *Transbound. Emerg. Dis.* 67, 1463–1471. doi: 10.1111/tbed.13515

Lu, C., Cai, Z., Zou, Y., Zhang, Z., Chen, W., Deng, L., et al. (2020). Flu phenotype - a one-stop platform for early warnings of the influenza a virus. *Bioinformatics*. 36, 3251–3253. doi: 10.1093/bioinformatics/btaa083

Lyons, D. M., and Lauring, A. S. (2018). Mutation and epistasis in influenza virus evolution. *Viruses* 10, 1–13. doi: 10.3390/v10080407

Mostafa, A., Abdelwhab, E. M., Mettenleiter, T. C., and Pleschka, S. (2018). Zoonotic potential of influenza a viruses: a comprehensive overview. *Virus. Basel* 10::497. doi: 10.3390/v10090497

Peacock, T. H. P., James, J., Sealy, J. E., and Iqbal, M. (2019). A global perspective on H9N2 avian influenza virus. *Viruses* 11:620. doi: 10.3390/v11070620

Peng, Y., Li, X., Zhou, H., Wu, A., Dong, L., Zhang, Y., et al. (2017). Continual antigenic diversification in China leads to global antigenic complexity of avian influenza H5N1 viruses. *Sci. Rep.* 7:43566. doi: 10.1038/srep43566

Peng, Y., Zhu, W., Feng, Z., Zhu, Z., Zhang, Z., Chen, Y., et al. (2020). Identification of genome-wide nucleotide sites associated with mammalian virulence in influenza a viruses. *Biosaf. Health* 2, 32–38. doi: 10.1016/j.bsheal.2020.02.006

Pu, J., Wang, S., Yin, Y., Zhang, G., Carter, R. A., Wang, J., et al. (2015). Evolution of the H9N2 influenza genotype that facilitated the genesis of the novel H7N9 virus. *Proc. Natl. Acad. Sci. U. S. A.* 112, 548–553. doi: 10.1073/pnas.1422456112

Quan, C., Shi, W., Yang, Y., Liu, X., Xu, W., Li, H., et al. (2018). New threats from H7N9 influenza virus: spread and evolution of high-and low-pathogenicity variants with high genomic diversity in wave five. *J. Virol.* 92:e00301-18. doi: 10.1128/JVI.00301-18

Su, S., Bi, Y., Wong, G., Gray, G. C., Gao, G. F., and Li, S. (2015). Epidemiology, evolution, and recent outbreaks of avian influenza virus in China. *J. Virol.* 89, 8671–8676. doi: 10.1128/JVI.01034-15

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., and Kawaoka, Y. (1992). Evolution and ecology of influenza a viruses. *Microbiol. Rev.* 56, 152–179. doi: 10.1128/mr.56.1.152-179.1992

Wu, A., Su, C., Wang, D., Peng, Y., Liu, M., Hua, S., et al. (2013). Sequential reassortments underlie diverse influenza H7N9 genotypes in China. *Cell Host Microb.* 14, 446–452. doi: 10.1016/j.chom.2013.09.001

Yu, H., Wu, J. T., Cowling, B. J., Liao, Q., Fang, V. J., Zhou, S., et al. (2014). Effect of closure of live poultry markets on poultry-to-person transmission of avian influenza a H7N9 virus: an ecological study. *Lancet* 383, 541–548. doi: 10.1016/S0140-6736(13)61904-2

Zhang, S.-Y., Song, S.-X., Liu, T., Sun, L., Wu, J.-L., He, Y.-J., et al. (2020). Epidemiological characteristics of human infection with avian influenza a(H7N9) virus in Shandong province from 2013 to 2017. *Modern Prevent. Med.* 47:7.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fmicb.2022.1079764

OPEN ACCESS

EDITED BY
Jianying Gu,
College of Staten Island,
United States

REVIEWED BY
Milad Zandi,
Tehran University of Medical Sciences,
Iran
Ruoyu Zhang,
Roche (China),
China

*CORRESPONDENCE
Yu-Dong Li
✉ lyd@zjsu.edu.cn
Ping-Ping Yao
✉ ppyaso@cdc.zj.cn
Jian-Min Jiang
✉ jmjiang@cdc.zj.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

RECEIVED 25 October 2022
ACCEPTED 15 December 2022
PUBLISHED 09 January 2023

CITATION
Sun Y-S, Sun H, Zhu H-P, Li G-L, Xu F, Lu H-J, Tang A, Wu B-B, Li Y-D, Yao P-P and Jiang J-M (2023) Comparative transcriptomic analyzes of human lung epithelial cells infected with wild-type SARS-CoV-2 and its variant with a 12-bp missing in the E gene.
*Front. Microbiol.* 13:1079764.
doi: 10.3389/fmicb.2022.1079764

# Comparative transcriptomic analyzes of human lung epithelial cells infected with wild-type SARS-CoV-2 and its variant with a 12-bp missing in the E gene

Yi-Sheng Sun[1†], Hao Sun[2†], Han-Ping Zhu[1], Gao-Lei Li[2], Fang Xu[1], Hang-Jing Lu[1], An Tang[3], Bei-Bei Wu[1], Yu-Dong Li[2]*, Ping-Ping Yao[1]* and Jian-Min Jiang[1]*

[1]Key Laboratory of Vaccine, Prevention and Control of Infectious Disease of Zhejiang Province, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China, [2]Department of Biological Engineering, School of Food Science and Biotechnology, Zhejiang Gongshang University, Hangzhou, China, [3]Key Laboratory of Health Risk Factors for Seafood of Zhejiang Province, Zhoushan Municipal Center for Disease Control and Prevention, Zhoushan, Zhejiang, China

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that caused a global outbreak of coronavirus disease 2019 (COVID-19) pandemic. To elucidate the mechanism of SARS-CoV-2 replication and immunogenicity, we performed a comparative transcriptome profile of mRNA and long non-coding RNAs (lncRNAs) in human lung epithelial cells infected with the SARS-CoV-2 wild-type strain (8X) and the variant with a 12-bp deletion in the E gene (F8). In total, 3,966 differentially expressed genes (DEGs) and 110 differentially expressed lncRNA (DE-lncRNA) candidates were identified. Of these, 94 DEGs and 32 DE-lncRNAs were found between samples infected with F8 and 8X. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyzes revealed that pathways such as the TNF signaling pathway and viral protein interaction with cytokine and cytokine receptor were involved. Furthermore, we constructed a lncRNA-protein-coding gene co-expression interaction network. The KEGG analysis of the co-expressed genes showed that these differentially expressed lncRNAs were enriched in pathways related to the immune response, which might explain the different replication and immunogenicity properties of the 8X and F8 strains. These results provide a useful resource for studying the pathogenesis of SARS-CoV-2 variants.

KEYWORDS

**COVID-19, RNA-seq, lncRNAs, E gene, immune response**

# Introduction

The novel coronavirus disease 2019 (COVID-19) is an acute respiratory infectious disease caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, which was first identified in China (Zhu et al., 2020). All the human beings are generally susceptible to the SARS-CoV-2 virus. By the 4th of December (Kim et al., 2022), it had led to more than 640 million patients and more than 6.6 million deaths worldwide (WHO, 2022). The SARS-CoV-2 virus is a single-stranded RNA virus and can mutate easily. Since the outbreak of COVID-19 in the late of 2019, the SARS-CoV-2 virus has continuously evolved with many variants emerging across the world, such as Alpha, Beta, Delta and Omicron variants (Cao et al., 2022; Iketani et al., 2022). In particular, the Omicron variants BA.4 and BA.5, containing the prominent immune escape characteristics, have spread all over the world and become the dominant strains currently (Cao et al., 2022). In China, a large part of the clustered outbreaks and sporadic cases were caused by the BA.5 infection (Feng et al., 2022; Jiang et al., 2022), which placed great pressure on COVID-19 prevention strategies. Therefore, it is necessary to pay more attention to SARS-CoV-2 variants.

The envelope (E) protein, located at the viral envelope, is the smallest one of four major structural proteins of SARS-CoV-2 virus. It consists of only 76 amino acids, but plays important roles in the viral life cycle, such as viral assembly, budding and so on (Castano-Rodriguez et al., 2018). A SARS-CoV strain lacking the E protein is attenuated *in vivo* (Regla-Nava et al., 2015). In our previous study, we isolated both the E gene wild-type SARS-CoV-2 strain 8X and the E gene mutant strain F8 from the same specimen (Sun et al., 2020). Although no significant difference in the viral titer and infectivity was found in the E gene mutant SARS-CoV-2 strain F8 and the E gene wild-type strain 8X, F8, which contains a 12-bp deletion in the E gene, could produce a higher S protein content and induce a quicker humoral immune response than 8X. The inactivated SARS-CoV-2 vaccine produced from the F8 strain could trigger higher levels of the IgG titer and neutralizing antibody titer than those from the 8X strain at 1 and 3 weeks post vaccination, respectively. It seemed that the E gene mutation could influence the replication and immunogenicity of SARS-CoV-2. However, the mechanism has not been elucidated yet.

In this study, the whole-transcriptome sequencing was performed based on the F8-and 8X-infected human lung epithelial (Calu-3) cells. Subsequently, differentially expressed genes (DEGs) and lncRNAs of the F8 and 8X groups were analyzed, followed by functional interaction prediction analysis. Although large amounts of data have proven that several lncRNAs are involved in different kinds of viral infections, the underlying mechanisms by which they act are still largely unknown. Based on the whole-transcriptome sequencing analysis, our results may provide novel insights into the molecular basis of SARS-CoV-2 infection.

# Materials and methods

## Ethics statement

All the experiments related to live SARS-CoV-2 viruses were approved by the Ethics Committee of the Zhejiang Provincial Center for Disease Control and Prevention (ZJCDC) in China, and carried out in the biosafety level 3 (BSL-3) laboratory of ZJCDC.

## Virus and cells

The SARS-CoV-2 clinical strains F8 and 8X were purified from the pharyngeal swab of a male COVID-19 patient in Hangzhou as mentioned previously (Yao et al., 2020). Calu-3 cells were obtained from the National Collection of Authenticated Cell Cultures and cultured in DMEM (Gibco, United States) with 10% fetal bovine serum (FBS, Every Green, China) at 37°C in a 5% $CO_2$ incubator. Cells were seeded into 6-well plates at a density of $1*10^6$ cells/well, and were infected by the F8 and 8X strains at a multiplicity of infection (MOI) of 2. Phosphate-buffered saline (PBS) was used as a negative control. Each group had two replicates. One hour post-adsorption at 37°C, the viral inocula were discarded, and the cells were maintained in the virus growth medium (DMEM containing 3% FBS) after washing twice with PBS. Two days post-infection, cells were collected, and the total RNA was extracted by using an RNeasy Plus Mini Kit.

## RNA extraction and next-generation sequencing

The RNA concentrations and quality of each sample were measured using Nanodrop One. The RNA integrity was detected by Agilent 2,100. Total RNA was used by depleting ribosomal RNA according to the manuscript of the Ribo-Zero rRNA Removal Kit. The rRNA-depleted RNA was fragmented, and the cDNA library was constructed using the TruSeq RNA sample Prep Kit (Illumina, San Diego, CA, USA). The sequencing libraries were sequenced on an Illumina NovaseqTM 6,000 platform according to the manufacturer's instructions. The sequence data generated from this project has been deposited in NCBI under SRA submission PRJNA909976.

## RNA-Seq data analysis

Transcript assembly: First, Cutadapt (Martin, 2011) was used to remove the reads that contained adaptor contamination, low-quality bases and undetermined bases. Then, sequence quality was verified using FastQC (Babraham Bioinformatics, 2022). We used Bowtie2 (Langmead and Salzberg, 2012) and Hisat2 (Kim et al., 2015) to map reads to the genome of Homo sapiens.

The mapped reads of each sample were assembled using StringTie (Pertea et al., 2015). Then, all transcriptomes from three samples were merged to reconstruct a comprehensive transcriptome using Perl scripts. After the final transcriptome was generated, StringTie and edgeR (Robinson et al., 2010) were used to estimate the expression levels of all transcripts.

LncRNA identification: First, transcripts that overlapped with known mRNAs and transcripts shorter than 200 bp were discarded. Then, we utilized CPC (Kong et al., 2007) and CNCI (Sun et al., 2013) to predict transcripts with coding potential. All transcripts with CPC scores $< -1$ and CNCI scores $<0$ were removed. The remaining transcripts were considered as lncRNAs.

Differential expression analysis of mRNAs and lncRNAs: StringTie was used to determine the expression level for mRNAs and lncRNAs by calculating FPKM (Trapnell et al., 2010). The differentially expressed mRNAs and lncRNAs were selected with log2 (fold change) $>1$ or log2 (fold change) $< -1$ and with statistical significance (I value $< 0.05$) by R package – edgeR (Robinson et al., 2010).

Target gene prediction and functional analysis of lncRNAs: To explore the function of lncRNAs, we predicted the cis-target genes of lncRNAs. LncRNAs may play a cis role by acting on neighboring target genes. In this study, coding genes in 100,000 upstream and downstream were selected by python script. Furthermore, trans-regulation analysis is a genome-wide search for well-associated target genes. Correlation analysis was performed between lncRNAs and the corresponding gene set. Associations with Pearson correlation coefficients greater than 0.4 ($p < 0.05$) were presumed to have targeted regulatory effects. Then, functional analysis of the target genes for lncRNAs was performed by using the BLAST2GO (Conesa et al., 2005).

## Go and KEGG enrichment analysis

Gene Ontology (GO) enrichment analysis of differentially expressed genes or lncRNA target genes was conducted with respect to biological process, molecular function, and cellular component. Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to perform pathway enrichment analysis. The R package clusterprofiler was used to perform the detailed enrichment analysis described above (Wu et al., 2021).

## Construction of the LncRNA-protein-coding gene co-expression network and competing endogenous RNA (ceRNA) network

For each lncRNA, the Pearson correlation coefficient of its expression value with that of each protein-coding gene was calculated. Under the conditions of an absolute value of the Pearson correlation coefficient $> 0.90$ and $p < 0.001$, the interaction network of the differentially expressed lncRNAs and protein-coding gene co-expression pairs was then constructed using Cytoscape (Shannon et al., 2003).

When constructing the competing endogenous RNA (ceRNA) network, lncRNAs were connected to the differentially expressed (DE) human miRNAs if they were predicted to interact with each other, and the upregulated (downregulated) miRNAs were connected to downregulated (upregulated) mRNAs if the former targeted the latter based on the database miRTarBase (Huang et al., 2020). The lncRNA-miRNA-mRNA network was visualized using Cytoscape (version 3.9.1). The co-expression network was built as follows: A custom database from a combination of public databases, starBase (version 2.0) and miRcode (version 11), was used to predict the interaction between known lncRNAs and miRNAs. The interaction of miRNA with novel lncRNA, circRNA, and mRNA was performed by using TargetScan (version 8.0) and miRanda (version 22.1), using default parameters. Pairs that appeared in both results were considered as prospective interactions. Subsequently, correlation analysis was performed between ceRNAs (circRNA, lncRNA, mRNA) and miRNAs. Based on the correlation between them, a hypergeometric distribution analysis was performed under the threshold of probability less than 0.05. The construction of the ceRNA network followed the following principles: (1) Pearson correlation coefficient within ceRNA pairs or between ceRNAs and miRNAs should be under the threshold of an absolute value of 0.4 ($p < 0.05$); and (2) At least one miRNA must satisfy the hypergeometric distribution test between two ceRNAs.

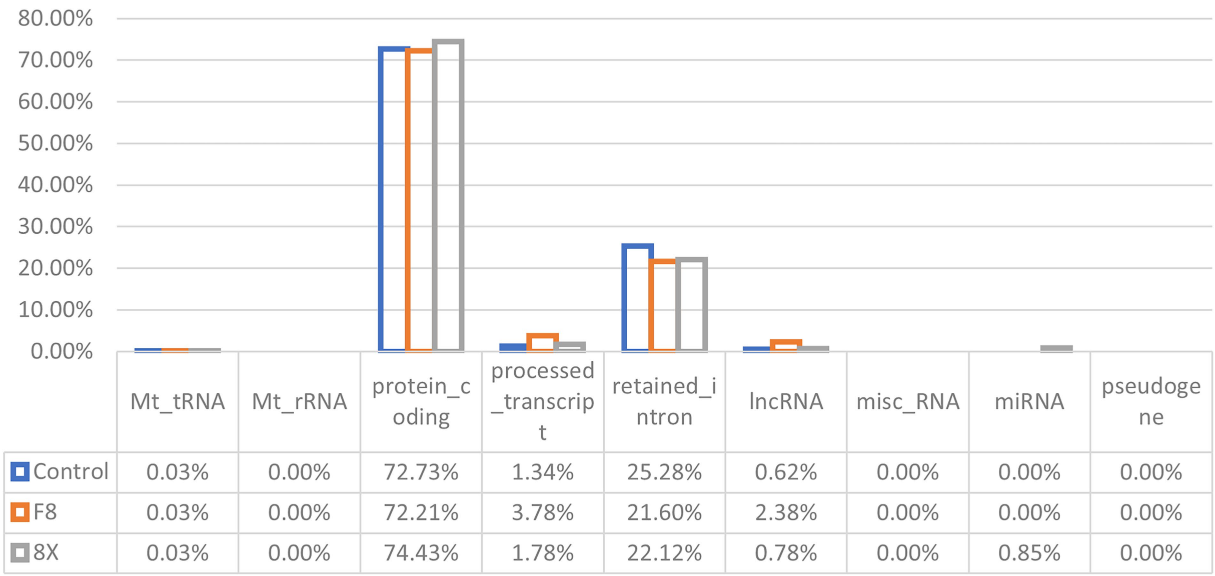
## Statistical analysis

All the statistical analyzes in this study were conducted in R (version 4.2.1). The Wilcoxon rank-sum test was used to compare the sample means between different groups, and was conducted with the wilcox.test function. Significance was expressed as a $p$ value $< 0.05$.

# Results

## Transcriptome profiles of SARS-CoV-2 infected cells

To identify different transcripts in SARS-CoV-2 infected cells, the transcriptomes of the Calu-3 cells infected with F8, 8X, or without SARS-CoV-2 infection as controls were detected using high-throughput RNA sequencing. Robust and reproducible data were obtained from all samples. After quality filtering, clean reads were mapped to the human reference genome (hg38) using HiSat2, and were assembled with StringTie. Then, coverage analysis was performed on these clean reads on different annotated gene types. The distribution of each type of gene was counted according to the expression level (Figure 1). In total, eight categories of RNA were identified according to database annotation of those transcripts, in



| | Mt_tRNA | Mt_rRNA | protein_coding | processed_transcript | retained_intron | lncRNA | misc_RNA | miRNA | pseudogene |
|---|---|---|---|---|---|---|---|---|---|
| 🟦 Control | 0.03% | 0.00% | 72.73% | 1.34% | 25.28% | 0.62% | 0.00% | 0.00% | 0.00% |
| 🟧 F8 | 0.03% | 0.00% | 72.21% | 3.78% | 21.60% | 2.38% | 0.00% | 0.00% | 0.00% |
| ⬜ 8X | 0.03% | 0.00% | 74.43% | 1.78% | 22.12% | 0.78% | 0.00% | 0.85% | 0.00% |

FIGURE 1
RNA categories of quantified transcripts in RNA-seq. The percentages of each RNA type are shown in F8/8X infected or mock-infected Calu-3 cells.

which the protein-coding genes were highly represented (72.21% in F8 and 74.43% in 8X, respectively).

## Identification of differentially expressed genes

DEGs were identified from the control vs. SARS-CoV-2 infection groups and used for functional annotation. All DEGs were demonstrated in volcano plots (Figures 2A–C). In the F8 vs. 8X group, there were 149 upregulated DEGs and 41 downregulated DEGs. The comparison between the F8 vs. control or 8X vs. control groups revealed that a total of 1,372 or 1,275 genes were significantly regulated after viral infection. A Venn diagram was plotted to identify the common or unique DEGs in human lung cells with or without SARS-CoV-2 infection (Figure 2D and Support file 2). A total of 14 common DEGs were identified. Given the exposure of cells to F8, the number of DEGs was smaller (345 genes vs. 394 genes) in comparison of the 8X vs. control group. The top 10 DEGs, including the E-selectin (SELE), colony stimulating factor 2 (CSF2) and pentraxin 3 (PTX3), were shown in the heatmap (Figure 2E).

## Go/KEGG analysis of DEGs

GO enrichment of the DEGs was divided into three types: biological process, cellular component, and molecular function (Figure 3A). Each top ten terms of the three types, such as cytokine activity and membrane raft, were displayed. The top 20 enriched pathways were shown through KEGG functional analysis of the DEGs (Figure 3B). Significant enrichment in the TNF signaling pathway, viral protein interaction with cytokine and cytokine receptor, and cytokine-cytokine receptor interaction were found.

## Identification of lncRNAs and Go/KEGG analysis of lncRNA target genes

The assembled transcripts were used to identify lncRNAs using the pipeline as described in the Methods. In total, 110 annotated and novel lncRNAs were identified. It has been reported that lncRNAs, in comparison with protein-coding genes, usually share some common genomic features with their sequences. However, they are generally shorter in length, have fewer but longer exons, and have lower expression levels and lower evolutionary sequence conservation (Liu et al., 2019). To further determine the characteristics of the lncRNAs identified in the present study, we compared transcript length, exon number and expression levels of protein-coding genes and lncRNAs. The results revealed that fewer exons and lower expression levels were also found in the lncRNAs, which was consistent with the reported lncRNAs (Figure 4).

We next performed GO and KEGG pathway analyzes for the target genes of lncRNAs between 8X and F8. The top 20 GO terms and KEGG pathways were reported in Figure 5. The GO term analysis divided differentially expressed lncRNAs into the same three types as DEGs. Biological processes such as response to chemokine and protein localization to cytoskeleton, the molecular functions such as cytokine activity and translation regulator activity, and the molecular functions such as membrane raft and transcription regulator complex, were all enriched. KEGG

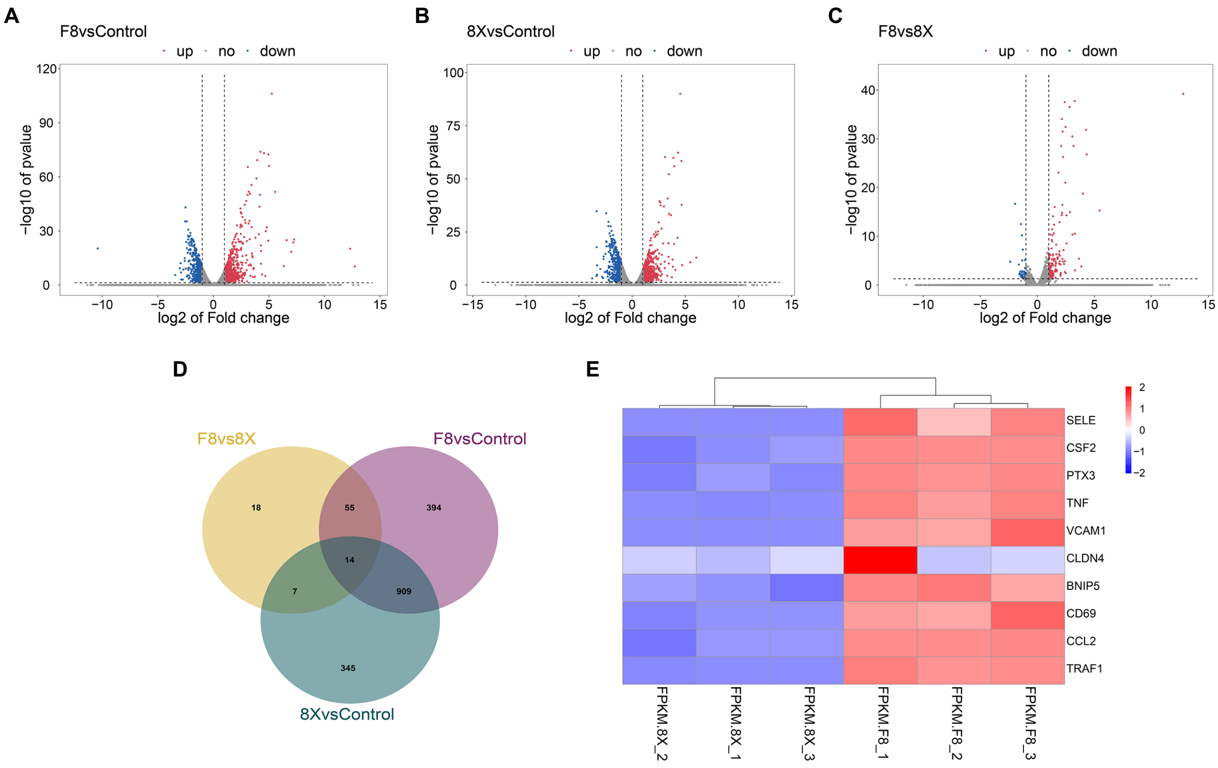


FIGURE 2
Volcano plots showing the DEGs between 8X and control **(A)**, F8 and control **(B)**, and the 8X and F8 **(C)**. **(D)** Venn diagram of the identified DEGs shared among the above three groups. **(E)** Heatmap of the top 10 DEGs between 8X and F8 groups.

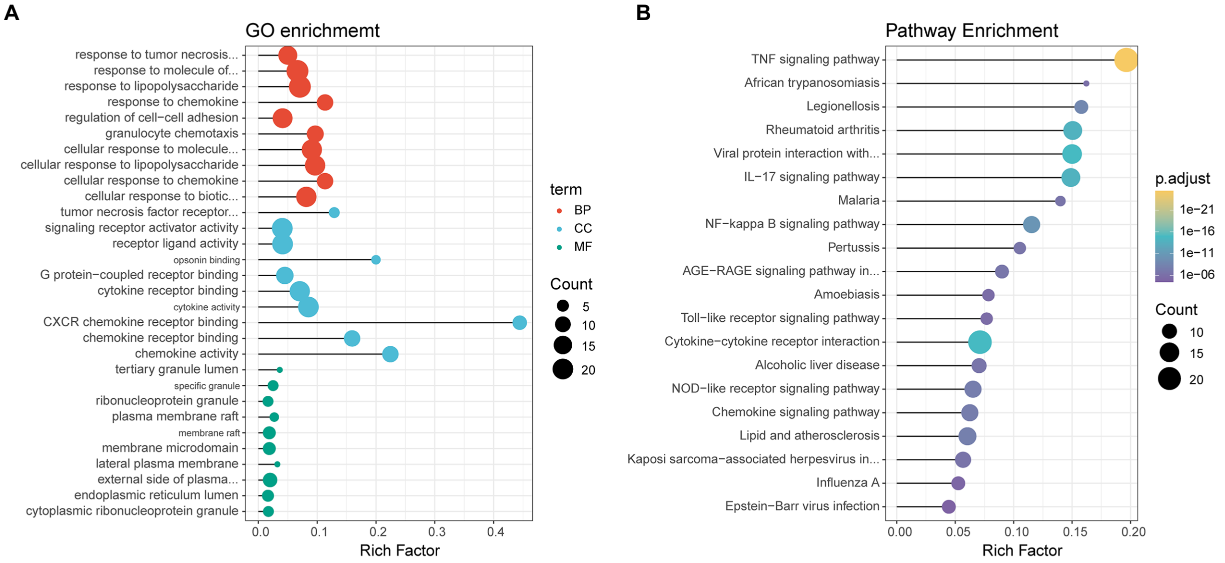


FIGURE 3
Dot plots displaying enriched GO terms **(A)** or KEGG pathways **(B)** of identified DEGs. Dot size represents the count of the enriched gene within each category, and the x-axis represents the gene ratio (rich factor). The GO terms/KEGG pathways are arranged on the basis of the value of the gene ratio, not on their adjusted *p* value, for easier visualization purposes.

enrichment analysis revealed that pathways related to the immune system, such as viral protein interaction with cytokine and cytokine receptor, chemokine signaling pathway and Toll-like receptor signaling pathways, were preferentially targeted. Therefore, the results of GO and KEGG pathway enrichment analyzes revealed that lncRNAs may act in cis or trans to

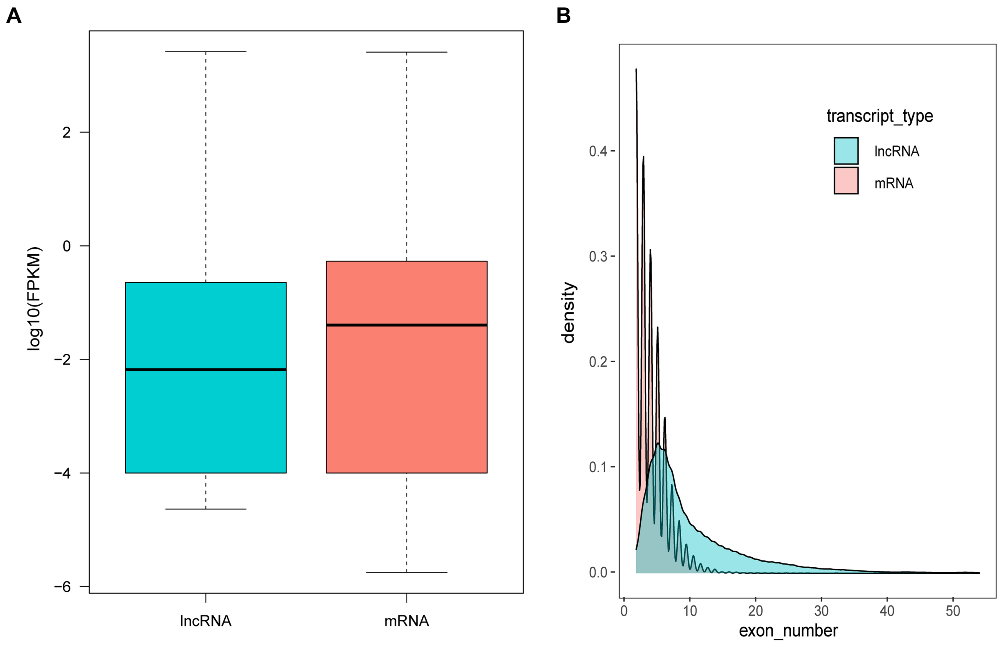


FIGURE 4
Comparison of the expression level **(A)** and exon number **(B)** of lncRNAs and protein-coding genes.

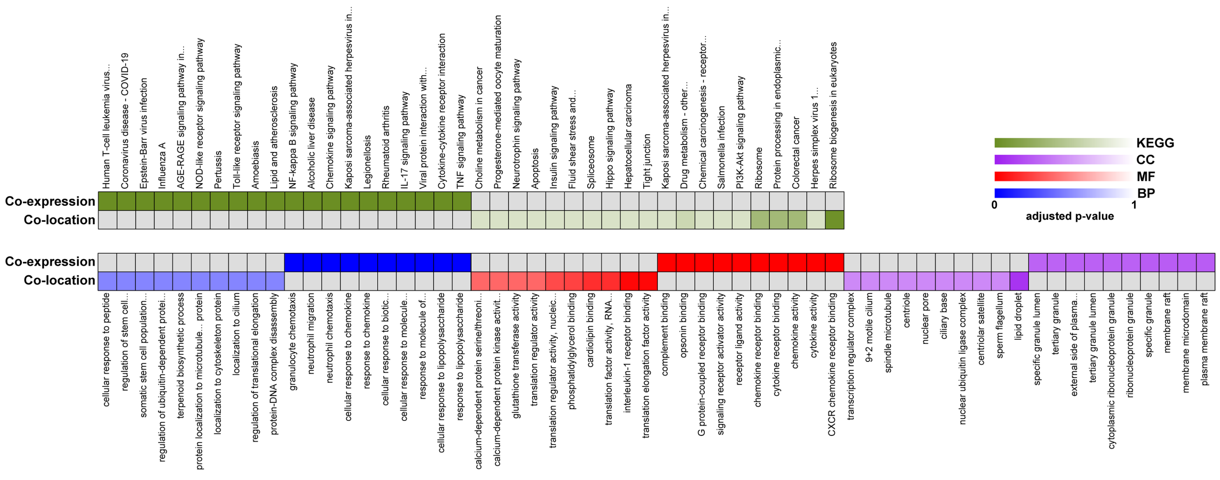


FIGURE 5
Functional enrichment analysis of identified lncRNAs between 8X and F8. Representative over-represented KEGG pathway (top) and GO terms (bottom) of gene-expression clusters. BP, biological process; MF, molecular function; CC, cellular component.

participate in the regulation of the expression of multiple important genes in different processes, including the immune response and protein localization.

## Expression correlation analysis

The functional association between regulatory lncRNAs and protein-coding gene transcripts can be determined by performing expression correlation analysis. To further investigate the potential mechanism of the SARS-CoV-2 associated lncRNAs, the DE lncRNAs and their predicted target DE protein-coding genes were investigated by delineating lncRNA-protein-coding gene functional interactions. Here we identified 325 pairs of DE lncRNA-DE protein-coding genes between 8X and F8 (Support file 1), containing 22 lncRNAs and 77 protein-coding genes ($p < 0.01$). The network of co-expressed

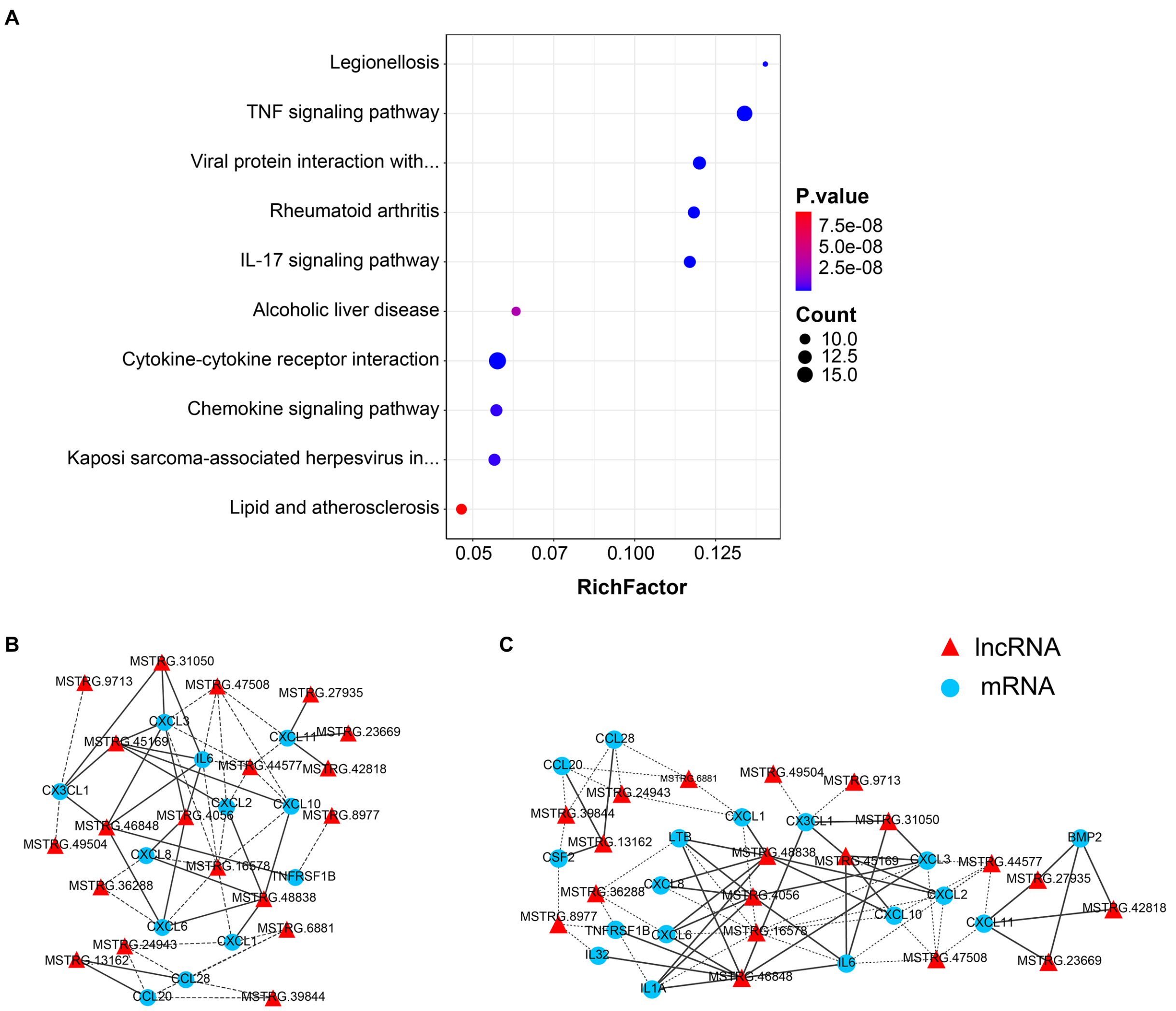


FIGURE 6

Co-expression analysis of lncRNAs and protein-coding genes. **(A)** The top 10 over-represented KEGG pathways of co-expressed genes. The network of lncRNAs with **(B)** viral protein interaction with cytokine and cytokine receptor, and **(C)** TNF signaling pathway-related genes ($p$<0.01). The solid line represents a positive correlation, while the dotted line represents a negative correlation.

lncRNA-protein-coding gene pairs based on a threshold of Pearson's correlation coefficient of 0.998. Next, the KEGG pathway analysis of co-expressed genes revealed the top 10 pathways, including the TNF signaling pathway, viral protein interaction with cytokine and cytokine receptor, and cytokine-cytokine receptor interaction, which were all significantly enriched (Figure 6A).

The network of lncRNAs with viral protein interaction with cytokine and cytokine receptor (Figure 6B), and TNF signaling pathway-related genes (Figure 6C) were analyzed. Several common cytokines, such as IL-6, CXCL2 and CXCL10, were found in all these pathways. Moreover, the CSF2, one of the top ten DEGs between 8X and F8 groups, was also involved in both the viral protein interaction with cytokine and cytokine receptor pathway and the TNF signaling pathway.

## CeRNAs network analysis

Based on the lncRNA-mRNA co-expression relationship and the regulation relationship of DE-miRNA-DE-mRNA and DE-miRNA-DE-lncRNA, the lncRNAs and mRNAs that were significantly differentially expressed and regulated by the same miRNA were screened. In total, 89 lncRNA-miRNA-mRNA interactions were finally obtained. Furthermore, according to the lncRNA-miRNA-mRNA and circRNA-miRNA-mRNA networks, differentially expressed circRNAs, lncRNAs, and mRNAs that were regulated by the same miRNA were further screened. Finally, 101 interaction pairs were obtained (Figure 7). Among these interactions, there were 2 upregulated lncRNAs and 2 upregulated mRNAs, and 1 upregulated miRNA in Calu-3 cells infected with the F8 variant compared with the 8X strain. There

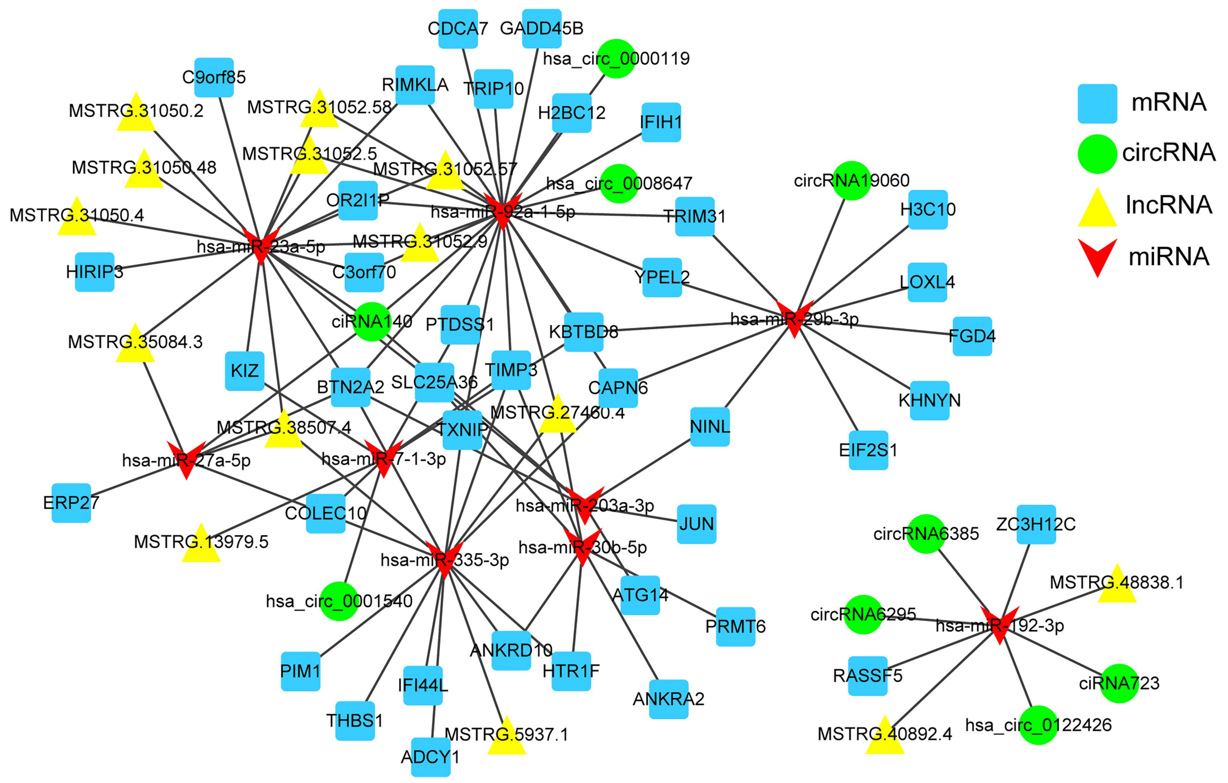


FIGURE 7
The competing endogenous RNA (ceRNA) network. mRNAs, circRNAs, host lncRNAs and miRNAs in the networks are represented as squares, circles, triangles and arrows, respectively.

were 12 lncRNAs, 25 upregulated and 11 downregulated mRNAs, and 1 upregulated and 7 downregulated miRNAs in Calu-3 cells infected with F8 or X8 compared with the control.

## Discussion

Since the outbreak of COVID-19, a large number of studies have focused on the spike protein of SARS-CoV-2. However, few studies have focused on the E gene. Both the E gene and S gene are the structural proteins of SARS-CoV-2. E gene is important for the virus maturation and replication (Regla-Nava et al., 2015). Studies related to the E gene were mainly about its use as a target for the virus detection (Kim et al., 2022; Valadan et al., 2022). It seems that the E gene of SARS-CoV-2 is conserved. However, in our previous study (Sun et al., 2020), we found that a 12-bp deletion in the E gene of the F8 variant was more likely generated as a result of viral adaptation to Vero cells. The E gene of SARS-CoV-2 might have several mutations. Since the E gene mutant strain F8 and E gene wild-type strain 8X were both from the same specimen, these two strains had similar genetic backgrounds. It would be interesting to use them to investigate the function of the E gene.

In this study, by applying transcriptome sequencing, we found 3,966 DEGs and 110 differentially expressed lncRNAs in Calu-3 cells infected with F8 compared with 8X. Direct functionly enrichment analysis of the DEGs showed that these genes were mainly involved in cytokines and chemokines, which are related to the host immune response. What's more, the co-expression analysis of differentially expressed lncRNAs and DEGs also showed the enrichment of the pathways involved in immune response, which may be critical for viral maturation, such as the TNF signaling pathway, and viral protein interaction with cytokine and cytokine receptor. A previous study showed that lncRNAs play a key role during viral infection (Liu et al., 2019). Additionally, some lncRNAs and miRNAs, such as hsa-miR-335-3p, hsa-miR-92a-1-5p, and hsa-miR-23a-5p, were identified as hub genes in the ceRNA network. Hsa-miR-92a-1-5p was found to enhance the interferon expression by targeting the suppressors of cytokine signaling 5 to inhibit feline panleukopenia virus replication in host cells (Liang et al., 2022). All the results strongly suggest that the genes involved in the immune response, especially the cytokines and chemokines, may play fundamental roles in the pathogenesis of different SARS-CoV-2 variants. Further study might be carried out in the SARS-CoV-2 E gene variant-infected animal models.

Through analysis of DEGs between the F8 and 8X groups, 85 upregulated genes were found. The E-selectin, CSF2 and PTX3 were the top 3 of the most upregulated genes. The E-selectin is important for leukocyte accumulation in inflammatory responses and PTX3 can amplify the immune response (Aslan et al., 2012; Ketter et al., 2016). In our previous research, a different immunogenicity of the inactivated COVID-19 vaccine produced by the F8 strain was found compared with that produced by the 8X strain (Sun et al., 2020). F8 vaccine might induce a higher expression level of E-selectin and

PTX3 than 8X vaccine to promote a quicker humoral immune response in mice. The TNF-α signaling pathway is also important for the immune response, and is one of the most upregulated pathways in SARS-CoV-2 infected A549-hACE2 cells (Sun et al., 2021). The release of TNF-α could trigger several cell adhesion molecules, such as selectins and VCAM-1, to induce inflammation *in vivo* (Kong et al., 2018), which is a protective biological response for eliminating viruses. In our study, the TNF signaling pathway was also enriched significantly through the KEGG functional analysis of DEGs, the co-expression correction analysis of regulatory lncRNAs, and protein-coding gene transcripts. The expression levels of TNF in the F8 group were 24.3-fold higher than those in the 8X group. It seems that the TNF signaling pathway is also related to the pathogenesis of different SARS-CoV-2 variants.

Cytokines and chemokines are important for the immune response (Berantini et al., 2022). Among the viral protein interaction pathway, the immune system pathway and the TNF signaling pathway, cytokines such as IL-6 and CSF2, and the chemokines such as CXCL2 and CXCL10, were all involved. CXCL2 was also found to be one of the common DEGs in SARS-CoV-2 infected Calu-3, hCM and A549-hACE2 (both low and high viral loads) cells (Sun et al., 2021). IL-6 might be a reliable indicator for the COVID-19 severity, and a higher level of IL-6 concentration was found in severe COVID-19 patients than in the moderate or mild groups (Bergantini et al., 2022). The CSF2 protein was inhibited to reduce the production of inflammatory factors and chemotaxis of inflammatory cells (Xiong et al., 2022). CXCL10, also known as interferon gamma-induced protein 10 (IP-10), is a pro-inflammatory chemokine (Coperchini et al., 2020). Higher expression levels of IL-6, CSF2 and CXCL10 were also found in the F8 group than in the 8X group, indicating a more severe inflammatory response in the immunized mice of the F8 group.

In conclusion, our study systematically characterized transcriptome profiles during SARS-CoV-2 infection and provides a comprehensive genome-wide resource for identifying and functionally analyzing the differentially expressed genes and non-coding RNAs. Immune response signaling, such as upregulation of IL-6, CSF2 and CXCL10 cytokines and a higher level of E-selectin and PTX3, as well as the TNF-α signaling pathway, might be the reason to explain the different pathogenesis caused by the E gene mutant strain F8 and E gene wild-type strain 8X. Further functional validations are needed to delineate the exact mechanistic details, and an animal model might be used to study the pathogenesis of both the E gene mutant and wild-type strains. These results will be useful for better understanding the pathogenesis of SARS-CoV-2 variants, and designing better preventive and therapeutic measures against viral infection.

## Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number: PRJNA909976.

## Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Zhejiang Provincial Center for Disease Control and Prevention (ZJCDC) in China. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

Y-SS, HS, P-PY, and Y-DL conceived and designed the experiments. Y-SS, H-PZ, FX, H-JL, and AT performed the experiments. G-LL, HS, and J-MJ analyzed the data. Y-SS, J-MJ, and Y-DL drafted the manuscript. All authors read and approved the final manuscript.

## Funding

This work was supported by the Basic Public Welfare Research Project of Zhejiang Province (LGF22C080002, LGF22H260024, LDT23H19013H19), the Key R&D Program of Zhejiang Province (2021C03197, 2021C03200), and the Zhejiang Provincial Foundation for Scientific Research in Medicine and Health (WKJ-ZJ-2220).

## Acknowledgments

We thank Jian-Hua Li from Zhejiang Provincial Center for Disease Control and Prevention for the technical assistance at the BSL-3 laboratory.

## 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.2022.1079764/full#supplementary-material

## References

Aslan, M. K., Boybeyi, O., Soyer, T., Senyücel, M. F., Ayva, S., Kısa, U., et al. (2012). Evaluation of omental inflammatory response with P−/E-selectin levels and histopathologic findings in experimental model. *J. Pediatr. Surg.* 47, 2050–2054. doi: 10.1016/j.jpedsurg.2012.06.024

Babraham Bioinformatics. (2022). *FastQC*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed July 20, 2022).

Bergantini, L., d'Alessandro, M., Cameli, P., Otranto, A., Luzzi, S., Bianchi, F., et al. (2022). Cytokine profiles in the detection of severe lung involvement in hospitalized patients with COVID-19: the IL-8/IL-32 axis. *Cytokine* 151:155804. doi: 10.1016/j.cyto.2022.155804

Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., et al. (2022). BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by omicron infection. *Nature* 608, 593–602. doi: 10.1038/s41586-022-04980-y

Castano-Rodriguez, C., Honrubia, J. M., Gutierrez-Alvarez, J., DeDiego, M. L., Nieto-Torres, J. L., Jimenez-Guardeno, J. M., et al. (2018). Role of severe acute respiratory syndrome coronavirus Viroporins E, 3a, and 8a in replication and pathogenesis. *MBio* 9, e02325–e02317. doi: 10.1128/mBio.02325-17

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Coperchini, F., Chiovato, L., Croce, L., Magri, F., and Rotondi, M. (2020). The cytokine storm in COVID-19: An overview of the involvement of the chemokine/chemokine-receptor system. *Cytokine Growth Factor Rev.* 53, 25–32. doi: 10.1016/j.cytogfr.2020.05.003

Feng, Z., Shen, Y., Li, S., Li, J., Wang, S., Zhang, Z., et al. (2022). The first outbreak of omicron subvariant BA.5.2 - Beijing municipality, China, July 4, 2022. *China CDC Wkly.* 4, 667–668. doi: 10.46234/ccdcw2022.136

Huang, H. Y., Lin, Y. C., Li, J., Huang, K. Y., Shrestha, S., Hong, H. C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 48, D148–d154. doi: 10.1093/nar/gkz896

Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J. F. W., Huang, Y., et al. (2022). Antibody evasion properties of SARS-CoV-2 omicron sublineages. *Nature* 604, 553–556. doi: 10.1038/s41586-022-04594-4

Jiang, H., Wu, C., Xu, W., Chen, H., Fang, F., Chen, M., et al. (2022). First imported case of SARS-CoV-2 omicron subvariant BA.5 - Shanghai municipality, China, may 13, 2022. *China CDC Wkly.* 4, 665–666. doi: 10.46234/ccdcw2022.104

Ketter, P., Yu, J. J., Cap, A. P., Forsthuber, T., and Arulanandam, B. (2016). Pentraxin 3: an immune modulator of infection and useful marker for disease severity assessment in sepsis. *Expert. Rev. Clin. Immunol.* 12, 501–507. doi: 10.1586/1744666x.2016.1166957

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kim, H. S., Lee, H., Park, J., Abbas, N., Kang, S., Hyun, H., et al. (2022). Collection and detection of SARS-CoV-2 in exhaled breath using face mask. *PLoS One* 17:e0270765. doi: 10.1371/journal.pone.0270765

Kong, D.-H., Kim, Y. K., Kim, M. R., Jang, J. H., and Lee, S. (2018). Emerging roles of vascular cell adhesion Molecule-1 (VCAM-1) in immunological disorders and cancer. *Int. J. Mol. Sci.* 19:1057. doi: 10.3390/ijms19041057

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

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

Liang, R., Liang, L., Zhao, J., Liu, W., Cui, S., Zhang, X., et al. (2022). SP1/miR-92a-1-5p/SOCS5: a novel regulatory axis in feline panleukopenia virus replication. *Vet. Microbiol.* 273:109549. doi: 10.1016/j.vetmic.2022.109549

Liu, J., Wang, F., Du, L., Li, J., Yu, T., Jin, Y., et al. (2019). Comprehensive genomic characterization analysis of lncRNAs in cells with Porcine Delta coronavirus infection. *Front. Microbiol.* 10:3036. doi: 10.3389/fmicb.2019.03036

Martin, M. (2011). CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17:10. doi: 10.14806/ej.17.1.200

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Regla-Nava, J. A., Nieto-Torres, J. L., Jimenez-Guardeño, J. M., Fernandez-Delgado, R., Fett, C., Castaño-Rodríguez, C., et al. (2015). Severe acute respiratory syndrome coronaviruses with mutations in the E protein are attenuated and promising vaccine candidates. *J. Virol.* 89, 3870–3887. doi: 10.1128/jvi.03566-14

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

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

Sun, G., Cui, Q., Garcia, G. Jr., Wang, C., Zhang, M., Arumugaswami, V., et al. (2021). Comparative transcriptomic analysis of SARS-CoV-2 infected cell model systems reveals differential innate immune responses. *Sci. Rep.* 11:17146. doi: 10.1038/s41598-021-96462-w

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41:e166. doi: 10.1093/nar/gkt646

Sun, Y.-S., Xu, F., An, Q., Chen, C., Yang, Z.-N., Lu, H.-J., et al. (2020). A SARS-CoV-2 variant with the 12-bp deletion at E gene. *Emerg. Microb. Infect.* 9, 2361–2367. doi: 10.1080/22221751.2020.1837017

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Valadan, R., Golchin, S., Alizadeh-Navaei, R., Haghshenas, M., Zargari, M., Mousavi, T., et al. (2022). Differential gene expression analysis of common target genes for the detection of SARS-CoV-2 using real time-PCR. *AMB Express* 12:112. doi: 10.1186/s13568-022-01454-2

WHO (2022). *WHO Coronavirus Disease (COVID-19) Dashboard*. Geneva, Switzerland: World Health Organization. Available at: https://covid19.who.int/. (Accessed December 4, 2022].

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* 2:100141. doi: 10.1016/j.xinn.2021.100141

Xiong, L., Cao, J., Yang, X., Chen, S., Wu, M., Wang, C., et al. (2022). Exploring the mechanism of action of Xuanfei Baidu granule (XFBD) in the treatment of COVID-19 based on molecular docking and molecular dynamics. *Front. Cell. Infect. Microbiol.* 12:965273. doi: 10.3389/fcimb.2022.965273

Yao, P., Zhang, Y., Sun, Y., Gu, Y., Xu, F., Su, B., et al. (2020). Isolation and growth characteristics of SARS-CoV-2 in Vero cell. *Virol. Sin.* 35, 348–350. doi: 10.1007/s12250-020-00241-2

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/fmicb.2023.1118025

OPEN ACCESS

EDITED BY
Zhan Zhou,
Zhejiang University,
China

REVIEWED BY
Yudong Li,
Zhejiang Gongshang University,
China
Zhao Zhang,
Fudan University,
China
Jianying Gu,
College of Staten Island,
United States

*CORRESPONDENCE
Shunyi Zhu
✉ Zhusy@ioz.ac.cn

SPECIALTY SECTION
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

RECEIVED 07 December 2022
ACCEPTED 01 February 2023
PUBLISHED 23 February 2023

CITATION
Gao B and Zhu S (2023) Mutation-driven parallel evolution in emergence of ACE2-utilizing sarbecoviruses.
*Front. Microbiol.* 14:1118025.
doi: 10.3389/fmicb.2023.1118025

# Mutation-driven parallel evolution in emergence of ACE2-utilizing sarbecoviruses

Bin Gao and Shunyi Zhu*

Group of Peptide Biology and Evolution, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Mutation and recombination are two major genetic mechanisms that drive the evolution of viruses. They both exert an interplay during virus evolution, in which mutations provide a first ancestral source of genetic diversity for subsequent recombination. Sarbecoviruses are a group of evolutionarily related β-coronaviruses including human severe acute respiratory syndrome coronavirus (SARS-CoV) and SARS-CoV-2 and a trove of related animal viruses called SARS-like CoVs (SL-CoVs). This group of members either use or not use angiotensin-converting enzyme 2 (ACE2) as their entry receptor, which has been linked to the properties of their spike protein receptor binding domains (RBDs). This raises an outstanding question regarding how ACE2 binding originated within sarbecoviruses. Using a combination of analyses of phylogenies, ancestral sequences, structures, functions and molecular dynamics, we provide evidence in favor of an evolutionary scenario, in which three distinct ancestral RBDs independently developed the ACE2 binding trait *via* parallel amino acid mutations. In this process, evolutionary intermediate RBDs might be firstly formed through loop extensions to offer key functional residues accompanying point mutations to remove energetically unfavorable interactions and to change the dynamics of the functional loops, all required for ACE2 binding. Subsequent optimization in the context of evolutionary intermediates led to the independent emergence of ACE2-binding RBDs in the SARS-CoV and SARS-CoV-2 clades of Asian origin and the clade comprising SL-CoVs of European and African descent. These findings will help enhance our understanding of mutation-driven evolution of sarbecoviruses in their early history.

KEYWORDS

**bat virus, SARS-CoV-2, insertion/deletion mutation, point mutation, evolutionary intermediate, functional diversification**

## Introduction

Coronaviruses (CoVs; *Coronaviridae*, *Nidovirales*) are a group of enveloped, single-stranded, positive-sense RNA viruses with a large RNA genome (~30 kb), comprising four genera (α-, β-, γ-, and δ; Nakagawa et al., 2016; Millet et al., 2021). The 5′-terminal two-thirds of their genomes contain two open reading frames (ORF1a and ORF1ab) coding for replicase polyproteins (pp1a and pp1ab) that are further processed into 16 nonstructural proteins (nsp; Nakagawa et al., 2016). The 3′-terminal one-third of the genome encode structural and accessory proteins. The structural proteins include spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, which are required for viral entry, assembly, trafficking, and release of virus particles (Siu et al., 2008; Li, 2016). Of the viral genome-encoding proteins, S protein is the most important determinant of viral infection in that it mediates viral attachment to specific host cell surface receptors and subsequent fusion and viral entry (Hulswit et al., 2016; Li, 2016; Piplani et al., 2021). This protein typically contains ~1,300 amino acids with some sites glycosylated. During viral entry, S protein is cleaved into two distinct structural and functional subunits (S1 and S2) at sites S1/S2 and S2' (Hulswit et al., 2016; Li, 2016; Piplani et al., 2021). S1 is composed of the N-terminal domain (NTD) and the C-terminal domain (CTD), both used as a receptor-binding domain (RBD) dependent on different viruses (Millet et al., 2021).

Severe acute respiratory syndrome coronavirus (SARS-CoV) and SARS-CoV-2 are two highly transmissible and pathogenic β-CoVs that caused serious pandemic in humans (Bolles et al., 2011; Cui et al., 2019; Arya et al., 2021; Harvey et al., 2021). They are two distantly related members of the *Sarbecovirus* subgenus (previously called lineage B) of the genus *β-Coronavirus*. Both viruses likely originated in bats, special reservoirs for emerging zoonotic pathogens (Dobson, 2005; Li et al., 2005a; Brook and Dobson, 2015; Cui et al., 2019; Boni et al., 2020). SARS-CoV and SARS-CoV-2 both use human angiotensin-converting enzyme 2 (ACE2; Kuhn et al., 2004), an enzyme involved in the regulation of cardiovascular and renal function *via* catalysis of angiotensin cleavage (Verano-Braga et al., 2020), as their entry receptor *via* the CTD of their spike protein known as RBD (Li F. et al., 2005; Li et al., 2005b; Hoffmann et al., 2020; Shang et al., 2020; Yan et al., 2020, 2021). The RBD structures of SARS-CoV and SARS-CoV-2 in complexed with human ACE2 (hACE2) have been solved with the aid of X-ray crystallography or cryo-electron microscopy (cryo-EM) techniques (Li F. et al., 2005; Shang et al., 2020). Their molecular cores are highly similar, both containing five anti-parallel β-strands (β1 to β4 and β7) and several short α-helices stabilized by three disulfide bridges (SS1 to SS3; Li F. et al., 2005; Shang et al., 2020). Three loops connect two core β-strands (β4 and β7) and are divided by two anti-parallel β-strands (β5 and β6). They protrude from the core scaffold to assemble a functional unit, named receptor-binding motif (RBM), responsible for direct interactions with hACE2 (Figure 1A). Accordingly, the three loops are, respectively, termed RBML1, RBML2, and RBML3, in which RBML2 is the longest one with one extra disulfide bridge (SS4). The RBM interacts with hACE2 through a large number of hydrophobic and hydrogen-bonding interactions (Figure 1B), in which the loops well match the shape of the highly exposed ACE2 helical regions (Li F. et al., 2005; Shang et al., 2020; Wang et al., 2020).

In addition to these two human viruses, some animal SARS-like CoVs (abbreviated as SL-CoVs) within the Sarbecovirus subgenus can also use ACE2 as their entry receptor, e.g., Rs4084, WIV1 and RaTG13 (Ge et al., 2013; Hu et al., 2017; Li et al., 2021). Their RBDs bind ACE2 with a similar mode to the two human viruses (Liu et al., 2021). Intriguingly, other SL-CoVs closely related to these ACE2-utilizing viruses do not use ACE2 as their receptor (Ren et al., 2008; Ge et al., 2013; Hu et al., 2017, 2018; Roelle et al., 2022). This raises an outstanding evolutionary question regarding how ACE2 binding originated within sarbecoviruses. One opinion thinks that ACE2 binding represents an ancestral and evolvable trait of sarbecoviruses and evolutionary deletions in two specific regions of RBDs led to the loss of the property in the ACE2 non-utilizing SL-CoVs (Shi and Wang, 2011; Starr et al., 2022); the other opinion insists that natural genetic recombination with other evolutionarily related viruses created this property (Boni et al., 2020; Wells et al., 2021). For example, based on phylogenetic reconciliation, it is inferred that extensive ancestral recombination might have occurred in sarbecoviruses including the SARS-CoV-2 lineage (Zaman et al., 2021). Comparative genomic analysis suggests that SARS-CoV-2 may have originated in the recombination of a virus similar to pangolin-CoV with one similar to RaTG13 (Xiao et al., 2020). However, Boni et al. proposed that SARS-CoV-2 itself is not a recombinant of any sarbecoviruses detected to date, and its receptor-binding motif could be an ancestral trait shared with bat viruses and not one acquired recently *via* recombination although the possibility of ancestral recombination events early in the evolution of sarbecoviruses is not excluded (Boni et al., 2020). In these studies, the authors' points of view are at opposite poles about the role of recombination in the evolution of SARS-CoV-2. Therefore, despite intensive studies worldwide (Hu et al., 2017; Cui et al., 2019; Boni et al., 2020; Xiao et al., 2020; Wells et al., 2021; Zaman et al., 2021), how these sarbecoviruses evolutionarily gained such ability especially in their early history is unresolved and certain to remain controversial, hindering a better understanding of their receptor shift to break through the species barrier.

Mutation and recombination are two major genetic mechanisms that drive the evolution of viruses *via* generating widespread molecular diversity. They both often exert an interplay during virus evolution, in which mutations provide a first ancestral source of functional diversity for subsequent recombination (Arenas et al., 2018). Therefore, although some studies have suggested the role of recombination in the evolutionary gain of ACE2 binding trait in some contemporary sarbecoviruses, it is very likely that mutations have driven the early origin of this trait among the phylogenetically distant ancestral species.

In this study, we employed a combination of analyses of phylogenies, ancestral sequences, structures, functions and molecular dynamics data of the sarbecovirus RBDs and found several key evolutionary events related to ACE2 binding, which had repeatedly occurred in the early evolution of all the three clades of this subgenus, including the SARS-CoV and SARS-CoV-2 clades of Asian origin and the clade comprising SL-CoVs of European and African descent. This suggests that their histories involve parallel evolution on distinct progenitors that ultimately gave rise to the ancestral ACE2-utilizing sarbecoviruses. The proposal of the possible existence of an evolutionary intermediate in the early history of Sarbecovirus evolution will help gain a better understanding of how the viruses gradually evolve to expand their entry mechanisms to enhance their fitness.

## Materials and methods

### LigPlot+ analysis of the RBD-hACE2 complex

For LigPlot+ analysis, hydrogen bonds and hydrophobic interactions were automatically calculated by the HBPLUS program (McDonald and Thornton, 1994; Laskowski and Swindells, 2011) where hydrogen-bond calculation parameters are 2.70 (maximum: H-A distance) to 3.35 (maximum D-A distance; here, H = hydrogen; A = acceptor; D = donor), and non-bonded contact parameters are 2.90 (minimum contact distance) to 3.90 (maximum contact distance). For hydrophobic contacts, hydrophobic atoms are carbon or sulfur. The treatment of connectivity records was used if possible (Laskowski and Swindells, 2011).

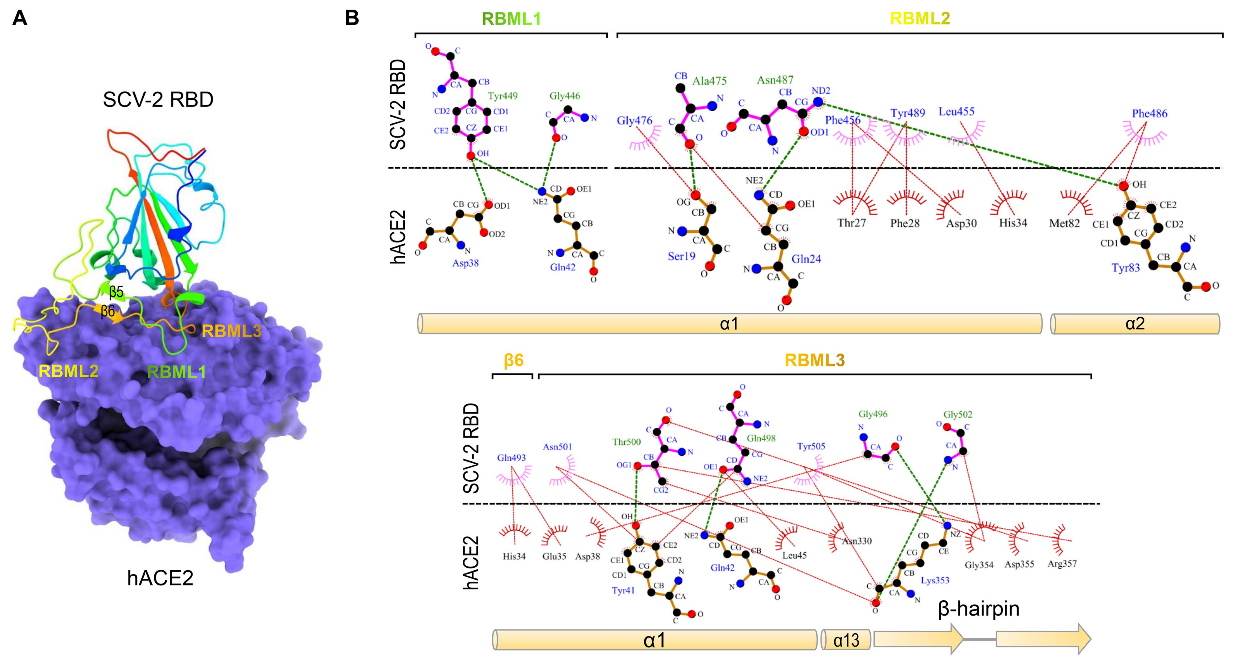


FIGURE 1
The SARS-CoV-2 RBD binds to hACE2 *via* residues located on the three loops. **(A)** The structure of SARS-CoV-2 RBD and hACE2 complex. The RBM comprising the three loops (designated as RBML1 to RBML3) docks onto the surface of hACE2 (shown in purple; pdb entry 6LZG). **(B)** LigPlot+ plot of the interaction diagram. Hydrophobic contacts and hydrogen bonding between the two loops (RBML1 and RBML2) of the RBD and the two α-helices (α1 and α2) of hACE2 are shown at the top and the interactions between RBML3 and α1, α13, and the β-hairpin of hACE2 at the bottom. The horizontal dotted line represents the interface, in which the residues involved in direct intermolecular hydrophobic contacts are shown as semicircles with radiating spoke and linked by red dotted lines and hydrogens (<4Å) are represented by green dashed lines.

## Construction of phylogenetic trees

For constructing the phylogenetic tree of RBDs from the *Sarbecovirus* subgenus, we firstly conducted BLASTP searching against the GenBank database[1] with SARS-CoV-2 RBD as query to collect homologs and then the retrieved sequences were aligned by ClustalX.[2] Using this alignment, we inferred a phylogenetic tree by the neighbor joining method with *p* distance to compute the evolutionary distances (NJp method) in the units of the number of amino acid differences per site with MEGA (Yoshida and Nei, 2016)[3]. As a comparison, we also inferred a tree using the Maximum Likelihood method with Whelan And Goldman (WAG) model and a discrete Gamma distribution to model evolutionary rate differences among sites with MEGA, which were chosen by the "Find Best DNA/Protein Model (ML)" mode with the lowest BIC scores (5226.99). Both methods generated similar results with good agreement. For constructing the phylogenetic tree of the whole genomes of the viruses, we conducted BLASTN searching the GenBank database using the full genome sequence of SARS-CoV-2 as query. The retrieved homologs (22 genomes belonging to *Sarbecovirus*; Supplementary Table 1) were aligned with ClustalW implemented in MEGA v10.1.7 (See footnote 3). Using the "Find Best DNA/Protein Model (ML)" model, we analyzed the aligned genome sequences to find the best model of nucleotide substitution for tree construction by maximum likelihood (ML) method. The best model obtained was GTR + G + I with the lowest BIC scores (364135.9), with which we constructed the tree with MEGA. To exam whether a non-*Sarbecovirus* outgroup has a potential impact on the topology of the tree and the evolution direction, we used Middle East respiratory syndrome coronavirus (MERS-CoV; Supplementary Table 1) as outgroup to reconstruct a rooted tree with the same method described above. To exclude the potential impact of RBDs on the whole genome-based tree, we built a sub-genome tree in which all the RBD-coding regions were deleted with the same method described here. The best model obtained was still GTR + G + I with the lowest BIC scores (327206.8). All these trees were built with 500 bootstrap replicates to provide confidence estimates for tree branches.

## Ancestral sequence reconstruction

FastML, a web server for probabilistic reconstruction of ancestral sequences (Ashkenazy et al., 2012), was used to reconstruct ancestral sequences of RBMs of representative sarbecoviruses. This method includes both joint and marginal reconstructions and is especially suitable for the sequences containing indel mutations since it integrates both indels and characters through indel-coding methodology to provide for each indel a presence ('1') or absence ('0') state in the input sequences. To this end, the amino acid sequences and the genome-based trees with or without the RBD-encoding region were chosen as input

[1] https://blast.ncbi.nlm.nih.gov/Blast.cgi
[2] http://www.clustal.org/
[3] https://www.megasoftware.net/

files. In this analysis, a discrete gamma distribution was used to account for rate variation among sites and four different evolutionary models of amino acid substitutions (JTT, LG, WAG, and Dayhoff) were chosen to best fits the data analyzed.

## Creation of RBD sequence logo

Two distinct subfamilies of RBDs divided by ClustalX (named RBD-L and RBD-S) were input into the Weblogo server[4] for creating sequence logos with default parameters. Using the two logos, we calculated the frequency for new amino acid emergence in the RBD-Ls relative to that in the RBD-Ss.

## Preparation of recombinant RBDs

The method for preparation of recombinant SARS-CoV-2 RBD through renaturation from *E. coli*-produced inclusion body (IB) has been reported previously (Gao and Zhu, 2021). According to this method, we produced recombinant proteins of BtRBD derived from the SL-CoV BtKY72 (Protein_id=APO40579.1, residues $N^{324} - P^{516}$) and its mutant BtRBD|GY with two residues (Gly-446 and Tyr-449) inserted in the RBML1. To this end, codon-optimized genes were synthesized from the Tsingke Biotechnology Co., Ltd. (Beijing, China) that were ligated into pET-28a(+) by *Nco* I and *Xho* I restriction enzyme sites with a His tag at both N- and C-termini. Recombinant plasmids were transformed into *E. coli* BL21(DE3) for auto-induction to accumulate IBs under the direction of the T7 promoter. The IBs were then renatured by the previously described method (Gao and Zhu, 2021). Further purification was carried out by size-exclusion chromatography (SEC) with a Superdex™ 75 Increase 10/300 GL column on an AKTA Pure 25 system (GE Healthcare Life Sciences, Pittsburgh, PA, United States) with 1xPBS, pH7.5 as the running buffer and a flow rate of 0.3 ml min$^{-1}$. Peak fractions were pooled and the samples were analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). Protein concentrations were determined by measuring the absorbance of the protein solution at 280 nm with a UV–VIS Spectrophotometer (NanoDrop2000). The sample was stored at-80°C for use. A Q-TOF mass-spectrometric method was used to determine molecular weights of the purified recombinant RBDs with HPLC-Q-TOF-MS (Agilent Technologies, Chandler, AZ, United States). Recombinant hACE2 (Gln18-Ser740) was purchased from KMD Bioscience (Tianjin, China) which was expressed in HEK293 cells with >95% purity.

## Surface plasmon resonance binding experiments

Surface plasmon resonance (SPR) was used to evaluate the binding of various RBDs to hACE2. The experiments were performed on a Biacore T100 instrument with a CM-5 sensor chip (GE Healthcare Life Sciences, United States) at 25°C according to the method previously described (Zhu et al., 2022). hACE2 was covalently linked on the CM5 sensor chip according to the amine coupling strategy (Nikolovska-Coleska, 2015). For pH scouting procedure, the running buffer used was 1xPBS-T, pH 7.5 with 0.05% Tween 20 and hACE2 was separately solubilized in 10 mM sodium acetate at a final concentration of 25 μg/ml with different pH 4.0, 4.5, and 5.0. For immobilization, the CM5 surface was first activated with two injections of 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC 0.4 M) and N-hydroxysuccinimide (NHS 0.1 M; v:v = 1:1) at a flow rate of 10 μl/min and then hACE2 solubilized in 10 mM sodium acetate, pH 4.5 at a final concentration of 25 μg/ml was injected. Non-reacted carboxylic groups on sensor chip surface was blocked by ethanolamine-HCl (1 M, pH 8.5) for 420 s at a flow rate of 10 μl/min. The final immobilization level was 1810 RU.

For detecting binding, an analyte (SARS-CoV-2 RBD, BtRBD or BtRBD|GY) was diluted with the running buffer PBS-T at indicated final concentrations. SARS-CoV-2 RBD was two-fold diluted to final concentrations of 1,000, 500, 250, 125, 62.5, 31.25, and 15.625 nM and BtRBD to final concentrations of 10, 5, 2.5, 1.25, and 0.625 μM. BtRBD-GY was four-fold diluted to final concentration of 40, 10, and 2.5 μM. Diluted samples were injected at a flow rate of 30 μl/min over the immobilized hACE2 during 60 s. Dissociation was monitored for 120 s by injecting the running buffer followed by additional washing for 180 s at a flow rate of 30 μl/min for the completely removal of specifically and non-specifically bound biological material from the surface. Responses were measured in RUs as the difference between active and reference channel. The binding curve was fitted with the software BIAevaluation v2.0.1 using 1:1 Langmuir binding model. The rational of using hACE2 to test the activity of BtRBD and its mutant BtRBD|GY was based on the work of Letko et al., in which the authors used hACE2 as the assay target to evaluate multiple bat-derived SL-CoVs with a long RBD (Letko et al., 2020). They found that many of them were able to use this human receptor for cellular entry (Letko et al., 2020). This experiment confirmed the functional conservation of ACE2 between human and bats, in support of the rational of our experiment.

## Molecular dynamics simulations

The structures for MD simulations included: (1) SARS-CoV-2 RBD (PDB entry 6LZG); (2) SARS-CoV-2 $\mathrm{RBD}_{\mathrm{woIN}}$; (3) SARS-CoV-2 $\mathrm{RBD}_{\mathrm{C21\_L3}}$; (4) SARS-CoV-2 $\mathrm{RBD}_{\mathrm{CtoS}}$ (Figure 2). The latter three structural models were built by comparative modelling with the DeepView Project Mode at the SWISS-MODEL server,[5] in which SARS-CoV-2 RBD was used as template. For each structure, a 20-ns MD simulations were performed with the GROMACS 2020.1 software package[6] using the OPLS (Optimized Potential for Liquid Simulations)-AA/L all-atom force field (2001 aminoacid dihedrals) and TIP3P model for explicit water. Solvent shell thickness was 1.5 nm for the monomers and 3.0 nm for the complex in a cubic box and the total charge of the simulated systems were neutralized by adding sodium or chloride ions. The detailed method has been described previously (Gao and Zhu, 2021). The root mean squared

[4] http://weblogo.berkeley.edu/logo.cgi

[5] https://www.expasy.org/

[6] https://www.gromacs.org/

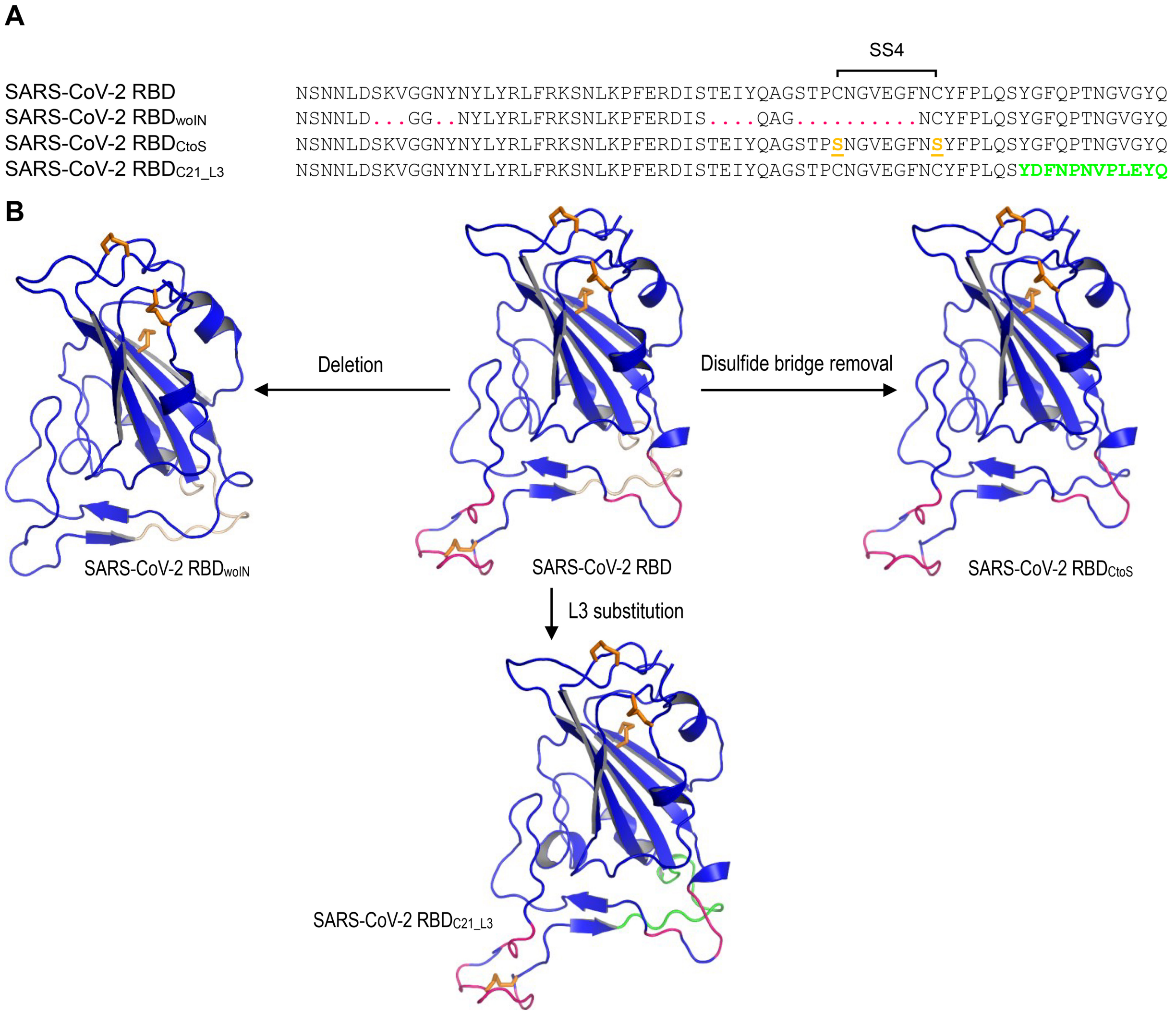

A
SS4
SARS-CoV-2 RBD NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ
SARS-CoV-2 RBDwoIN NSNNLD...GG..NYLYRLFRKSNLKPFERDIS....QAG.........NCYFPLQSYGFQPTNGVGYQ
SARS-CoV-2 RBDCtoS NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPSNGVEGFNSYFPLQSYGFQPTNGVGYQ
SARS-CoV-2 RBDC21_L3 NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYDFNPNVPLEYQ
B
Deletion
Disulfide bridge removal
SARS-CoV-2 RBDwoIN
SARS-CoV-2 RBD
SARS-CoV-2 RBDCtoS
L3 substitution
SARS-CoV-2 RBDC21_L3

FIGURE 2
The SARS-CoV-2 RBB mutants for MD simulations. **(A)** The designed RBD mutant sequences. SARS-CoV-2 RBD$_{woIN}$ represents a deletion mutant with corresponding amino acids in RBML1 and RBML2 (marked in red) deleted and "woIN" denotes "without insertions." SARS-CoV-2 RBD$_{CtoS}$ represents a mutant of two Cys to Ser mutations that remove the SS4 (underlined once and shown in orange). SARS-CoV-2 RBD$_{C21\_L3}$ represents a mutant whose RBML3 is substituted by the equivalent of the CoVZXC21 RBD (marked in green). **(B)** Structures of the SARS-CoV-2 RBD mutants generated by homology modelling with SARS-CoV-2 RBD (PDB entry 6LZG) as template.

deviation (RMSD) for measuring the difference of simulated structures to the structure present in the minimized, equilibrated system, and Cα root-mean-square-fluctuation (RMSF) that captures the fluctuation for each atom about its average position and gives insight into the flexibility of different structural regions of the simulated protein were calculated with the gmx rms command of GROMACS. In addition, for evaluating the lifetime of the three hydrogen bonds between SARS-CoV-2 RBML1 and hACE2, a 100-ns MD simulations were performed with the method described above except the solvent shell thickness of 3.0 nm instead of 1.5 nm.

## Statistics

Data in Supplementary Figure 1 are expressed as mean ± standard deviation (SD; $n = 2{,}001$) and statistical significance of means between two groups was determined by one-way analysis of variance (ANOVA) with SPSS Statistics 17.0 (SPSS Inc.).

# Results

## Mutation-driven evolution of RBDs in sarbecoviruses

Previous studies have found that some bat SL-CoVs with shorter RBML1 and RBML2 in their RBDs are unable to use ACE2 as their entry receptor (Hu et al., 2017), pointing out that the loop length evolution may be related to functional diversification between sarbecoviruse RBDs. To establish a correlation between the loop length and ACE2 binding, we systematically studied a group of RBDs from SARS-CoV, SARS-CoV-2 and SL-CoVs from bats and palm civets, which contained members with both short and long RBMLs

(Appendix 1). The mutations considered here included insertion/deletions (indels) altering loop sizes and point mutations altering amino acid sequences. For the uncharacterized RBDs, we used a phylogenomics method to correlate their sequences to the function (ACE2 binding or not). This method overlays known functions onto a phylogenetic tree, on which a sequence's function can be assigned by its phylogenetic position relative to the characterized ones (Eisen, 1998). To this end, we built a neighbor-joining (NJ) tree based on the amino acid sequences of RBDs (Figure 3A), from which two distinct structural subfamilies were clearly assigned. The ML method yielded a similar tree (Supplementary Figure 2). We named the long RBDs RBD-L and the short ones RBD-S. For the subfamily RBD-L, all members have two extended RBMLs (i.e., RBML1 and RBML2) in length with a 13–18 residues of extension relative to the members from the subfamily RBD-S (Drexler et al., 2010; Tao and Tong, 2019; Letko et al., 2020).

Phylogenomics analyses showed that the tree partitions were overall correlated with the RBD length and their functional properties, in which all members in the RBD-L subfamily are able to bind ACE2 (Ren et al., 2008; Hu et al., 2017, 2018; Letko et al., 2020; Roelle et al., 2022; Starr et al., 2022) except two bat SL-CoVs isolated from Europe (namely BatCoV-BM48-31-BGR-2008, abbreviated as BM48-31, and BatCoV-BB9904-BGR-2008, abbreviated as BB9904; Drexler et al., 2010). The functional loop length of these RBDs is slightly shorter than that of other RBD-Ls (Appendix 1) and the inability of binding to ACE2 has been experimentally confirmed recently in BM48-31 (Letko et al., 2020) and BB9904 (Roelle et al., 2022). For the RBD-S subfamily, all the members are unable to bind ACE2 (Ren et al., 2008; Hu et al., 2017, 2018; Letko et al., 2020; Roelle et al., 2022; Starr et al., 2022; Figure 3A), suggesting that they use an unidentified non-ACE2 receptor in mammals. The overall consistency among the loop indel pattern, the RBD tree topology and functional classification highlights the role of indels in the evolution of ACE2 binding within sarbecoviruses.

We subsequently conducted evolutionary conservation analyses to identify subfamily-specific amino acid positions (Figure 3B). It was found that seven strictly defined positions showed identity within one subfamily but difference in another (Figure 3B), indicating that they are a class of tree determinants that are likely relevant to functional

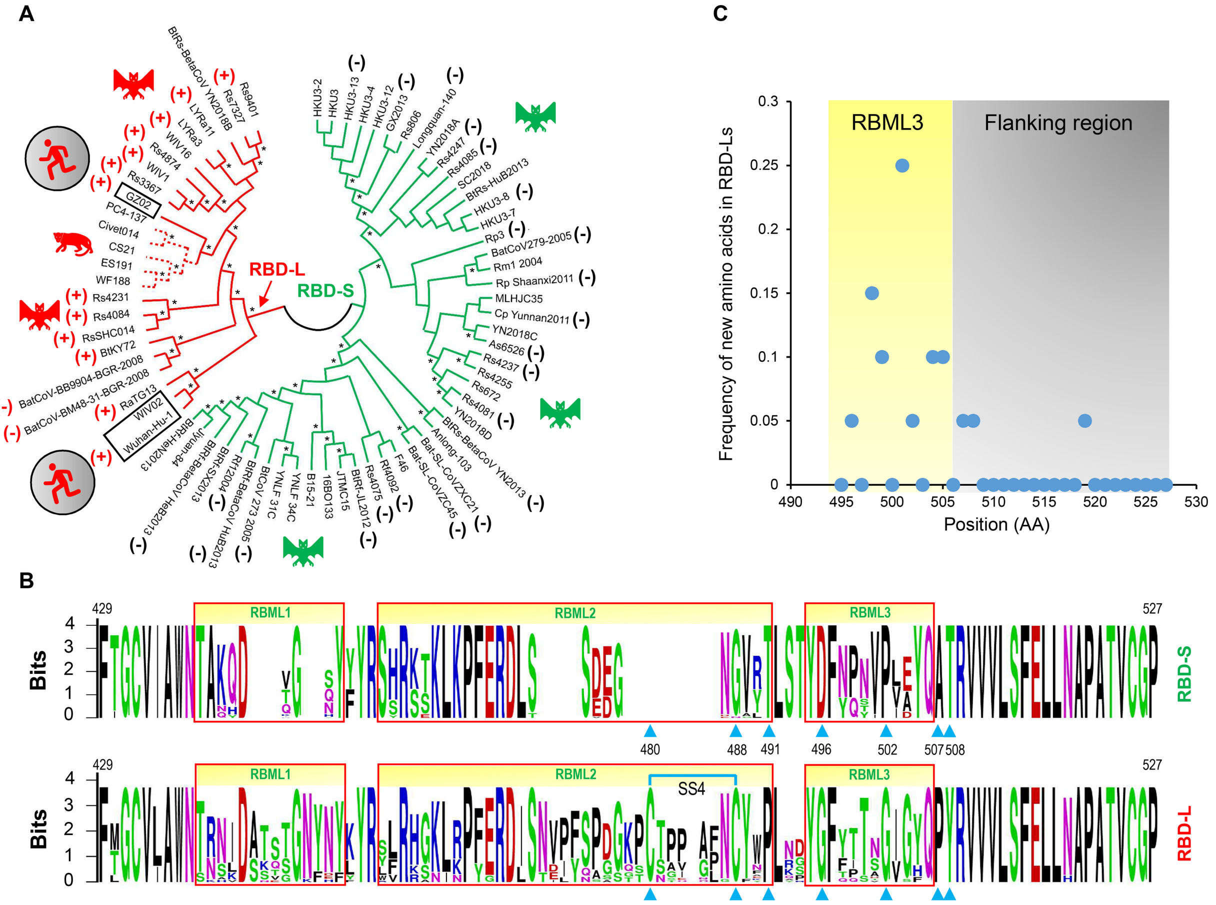


FIGURE 3
The classification and evolutionary conservation of *Sarbecovirus* RBDs. **(A)** The phylogeny-based classification. The tree was constructed by MEGA with the NJ method based on *p*-distance model of amino acid substitutions (NJp). Asterisks shown at nodes indicate the braches supported by up to 50% bootstrap based on 500 replicates. Two subfamilies differentiated by the tree are denoted as RBD-L (shown in red) and RBD-S (shown in green) based on their loop length. RBDs known to use ACE2 are marked by plus signs and those incapable of binding ACE2 by minus signs (data derived from Ren et al., 2008; Hu et al., 2017, 2018; Letko et al., 2020; Roelle et al., 2022; Starr et al., 2022). Hosts of the viruses, including bats, palm civet and human, are shown here. **(B)** The evolutionary conservation of RBDs analyzed by "Weblogo." The three loops are boxed in red and positions as tree determinants are indicated by cyan triangles. Top: RBD-S. Bottom: RBD-L. The SS4 in RBD-L is indicated by cyan lines. **(C)** Comparison of the frequency of new amino acids emergence in RBML3 and its flanking region.

diversification (Valencia and Pazos, 2003). These included Cys-480 and Cys-488 (both forming the SS4), Pro-491, Gly-496, Gly-502, Pro-507, and Tyr-508 in the RBD-L subfamily and the equivalent residues in the RBD-S subfamily are a residue deficiency at 480, Gly-488, Thr-491, Asp-496, Pro-502, Ala-507, and Thr-508 (numbering according to the SARS-CoV-2 RBD; Figure 3B). Because prior studies have shown that in eukaryotic genomes indel mutations often induce an increase in the substitution rate of their flanking regions (Tian et al., 2008; Zhang et al., 2011), we analyzed the frequency of the emergence of new amino acids in the RBML3 of the RBD-L subfamily compared with that of the RBD-S subfamily. The result showed that the RBML3 had a substitution rate of 0.05–0.25 calculated from 20 natural amino acids, which was far higher than that of its flanking region (Figure 3C). These observations suggest that loop extension, tree determining-related point mutations and accelerated substitutions in RBML3 commonly contribute the emergence of ACE2 binding in an ancestral RBD scaffold.

## Structural and functional significance of mutations

To study the potential effects of loop extension and amino acid substitutions on the dynamics of ACE2-binding RBDs, we designed three mutants of the SARS-CoV-2 RBD (Figure 2) for molecular dynamics (MD) simulations. They included: (1) SARS-CoV-2 $\text{RBD}_{\text{woIN}}$ with RBML1 and RBML2 extensions deleted; (2) SARS-CoV-2 $\text{RBD}_{\text{CtoS}}$ with two Cys to Ser mutations to remove the SS4 in RBML2; (3) SARS-CoV-2 $\text{RBD}_{\text{C21\_L3}}$ with the RBML3 substituted by the equivalent of the RBD from CoVZXC21, a member belonging to the RBD-S subfamily (Figure 3A). A 20-ns MD simulations revealed that the SARS-CoV-2 RBD exhibited a lower structural stability than the $\text{RBD}_{\text{woIN}}$, as identified by their RMSD values (~3.0 vs. 2.0 Å) for backbone atoms when calculated in an equilibrium state (15–20 ns; Figure 4A, left). Consistently, the wild-type RBD had a gyration radius of ~18.5 Å greater than that of $\text{RBD}_{\text{woIN}}$ (~16.9 Å; Figure 4A, right). These data show that the loop extensions in an ancestral RBD incapable of binding ACE2 caused a decrease in the stability of the new molecule but accompanying the emergence of a novel function, indicative of a structure–function trade-off in the RBD evolution, as observed in the evolution of some enzymes, in which they obtained new enzymatic specificities but accompanying the loss of the protein's stability (Shoichet et al., 1995; Tokuriki et al., 2008).

To examine the effects of the SS4 mutation and the RBML3 substitution (Figure 3C) on the flexibility of different structural regions of RBDs, we calculated their RMSFs for each simulated RBD structures based on the Cα atoms to study the fluctuation degree of the individual amino acids during simulations. By background subtraction of the wild-type RBD RMSF, we found that these two mutations primarily influenced the local flexibility of RBML2 (Figure 4B, left). Consistently, a "sausage" model analysis of the simulated structures showed that this loop in SARS-CoV-2 RBD exhibited greater structural flexibility than that of the two mutants (i.e., $\text{RBD}_{\text{C21\_L3}}$ and $\text{RBD}_{\text{CtoS}}$; Figure 4B, right). These data suggest that the conformation of RBML2 might be allosterically regulated by mutations at RBML3 (Figures 3C, 4B) in a distant manner or by the evolution of one new disulfide bridge (SS4) in its own region. The former well explains the cause of accelerated substitutions in RBML3 (Figure 3C) when evolved into an ACE2-binding RBD. For the latter, although the prevailing view is that disulfide bridges have been added during evolution to enhance the stability of proteins (Hogg, 2003), it appears that the added SS4 works as a regulator for the conformational flexibility of RBML2.

To study the functional role of loop extensions in ACE2 binding, we compared the dynamics of each loop between the *apo*- (receptor-free system) and ACE2-bound conditions. The time-curves of RMSDs during simulations showed that the RBML1 remained stable in both *apo* and ACE-bound conditions whereas ACE2 binding slightly stabilized the structure of RBML3 (Supplementary Figure 1). Remarkably, RBML2 exhibited a highly conformational flexibility in its *apo* state but ACE binding reduced the flexibility (Figure 4C, left). From the simulation trajectories, we extracted two distinct conformational states (herein named open and closed), in which only the open one is suitable for ACE binding (Figure 4C, middle). Such conformational flexibility may be mediated by Pro-491 because reverse mutation (Pro491Thr) can significantly decrease the flexibility of this loop in the SARS-CoV-2 RBD (Figure 4C, right). Therefore, the location of a proline on the last position of RBML2 (Figure 3) likely acts as a backbone switch controlled by prolyl *cis-trans* isomerization, which allows it to adopt two completely distinct conformations (*cis* and *trans*), as previously documented in other proteins (Schmidpeter and Schmid, 2015).

Among the seven tree determinants recognized here, threes (Cys-480, Cys-488 and Pro-491) have been found to play a potential role in conferring ACE2 binding *via* conformational modulation. Further structural analysis highlights the evolutionary significance of two other tree determinants (D496G and P502G). According to the determined structures of ACE2 complexed with SARS-CoV or SARS-CoV-2 RBD (Li F. et al., 2005; Wang et al., 2020), it can be proposed that the RBD-Ss are energetically unfavorable for ACE2 binding since there exist the electric charge repulse between Asp-496 of these RBDs and Asp-38 of ACE2 and the steric hindrance between Pro-502 of the RBDs and Lys-353 of ACE2 (Figure 4D). Substitutions by introducing a small glycine at these two positions (D496G and P502G) remove the energetically unfavorable interactions and create new H-bonds in the interface (Figures 1B, 4D).

## Phylogenetic evidence for ancestral parallel evolution

To infer the ancestral state of the mutations related to functional diversification, we reconstructed a phylogenetic tree based on the whole genome sequences of SARS-CoV, SARS-CoV-2 and related SL-CoVs (Figure 5), which is similar to a tree previously published (Lu et al., 2020). We found that adding a non-*Sarbecovirus* outgroup did not substantially alter the topology and the evolution direction of the tree (Supplementary Figure 3). This genome tree is topologically divided into three well supported clades (Figure 5). Clade 1 includes two bat SL-CoVs from Bulgaria and Kenya; clade 2 comprises SARS-CoV-2 and its bat relatives; and clade 3 contains SARS-CoV and its bat relatives. Different from the RBD tree, clades 2 and 3 in this genome tree show no correlation to the indel pattern described above rather than a mixed form of long and short RBDs (Figures 3A, 5). Given new evidence in support of ACE2 binding as an ancestral trait of sarbecoviruses, there are two competitive hypotheses that can

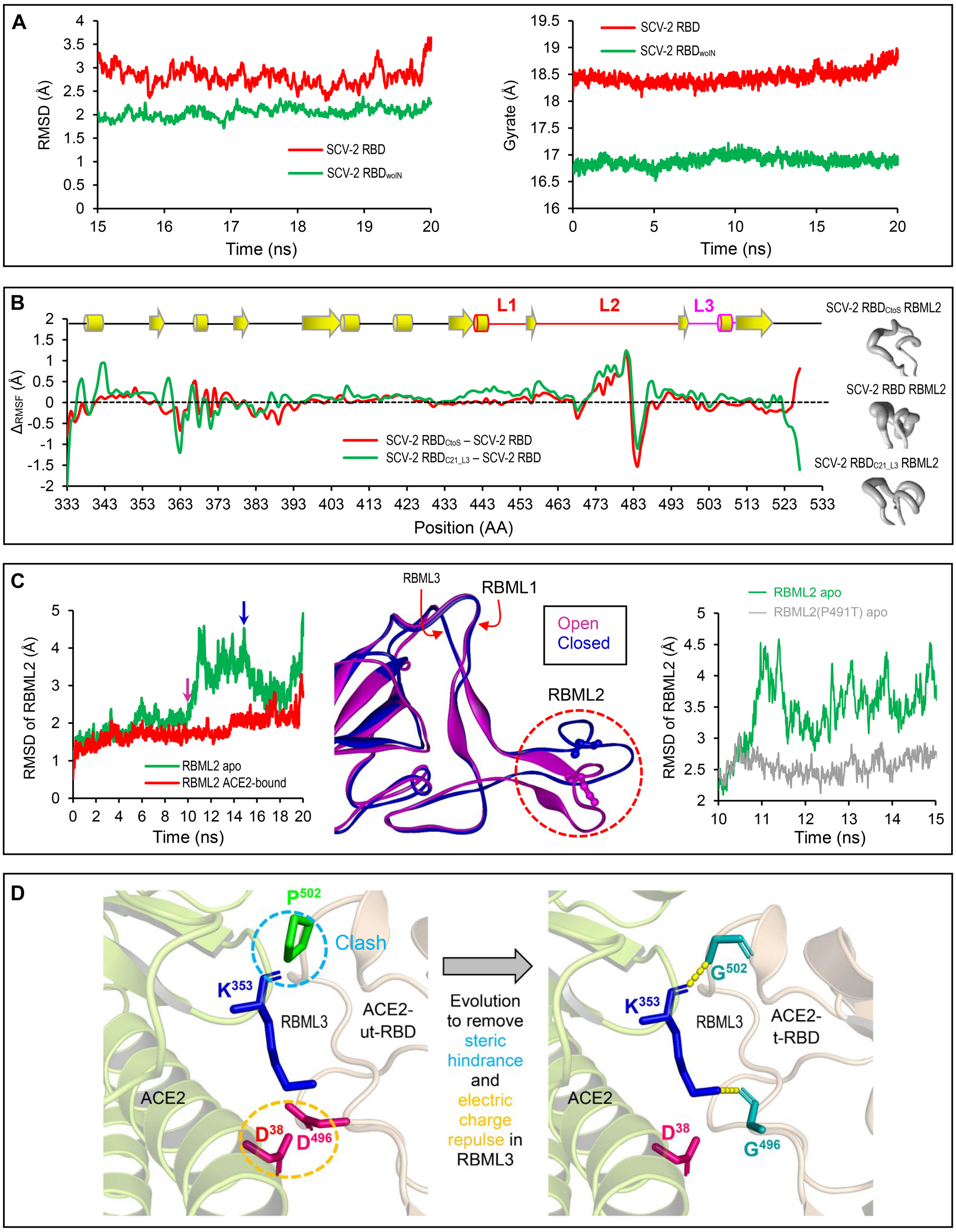


FIGURE 4
Structural and dynamics evidences for ACE2 binding origin. **(A)** Backbone-RMSDs of SARS-CoV-2 RBD and its deletion mutant shown as a function of time (left). Gyrate of proteins. SARS-CoV-2 RBD and its deletion mutant shown as a function of time (right). **(B)** ΔCα-RMSF data. SARS-CoV-2 RBDCtoS − SARS-CoV-2 RBD is marked in red and SARS-CoV-2 RBDC21_L3 − SARS-CoV-2 RBD in green (left). Conformational ensembles of RBML2 generated by MD simulations and shown by a "sausage" model with MOLMOL (right). **(C)** A 20-ns MD simulations showing structural dynamics of RBML2 in the apo state or ACE2-bound sate (left). Snapshots extracted from the MD trajectories at 10 and 15ns, respectively, showing two distinct conformations in RBML2 (open and closed; middle). Comparison of the RBML2 RMSDs between SARS-CoV-2 RBD and the P491T mutant (right). **(D)** Structural mapping showing parallel molecular evolution removing steric hindrance and electric charge repulse present in the ancestral state. The clash occurs between Pro-502 of RBDs incapable of binding ACE2 and Lys-353 of ACE2, indicated by a cyan dashed circle, and the electric charge repulse between Asp-496 of the RBDs incapable of binding ACE2 and Asp-496 of ACE2, indicated by an orange dashed circle. In the RBD-hACE2 complex, hydrogen bonds are shown by yellow dashed lines and involved residues displayed as sticks.

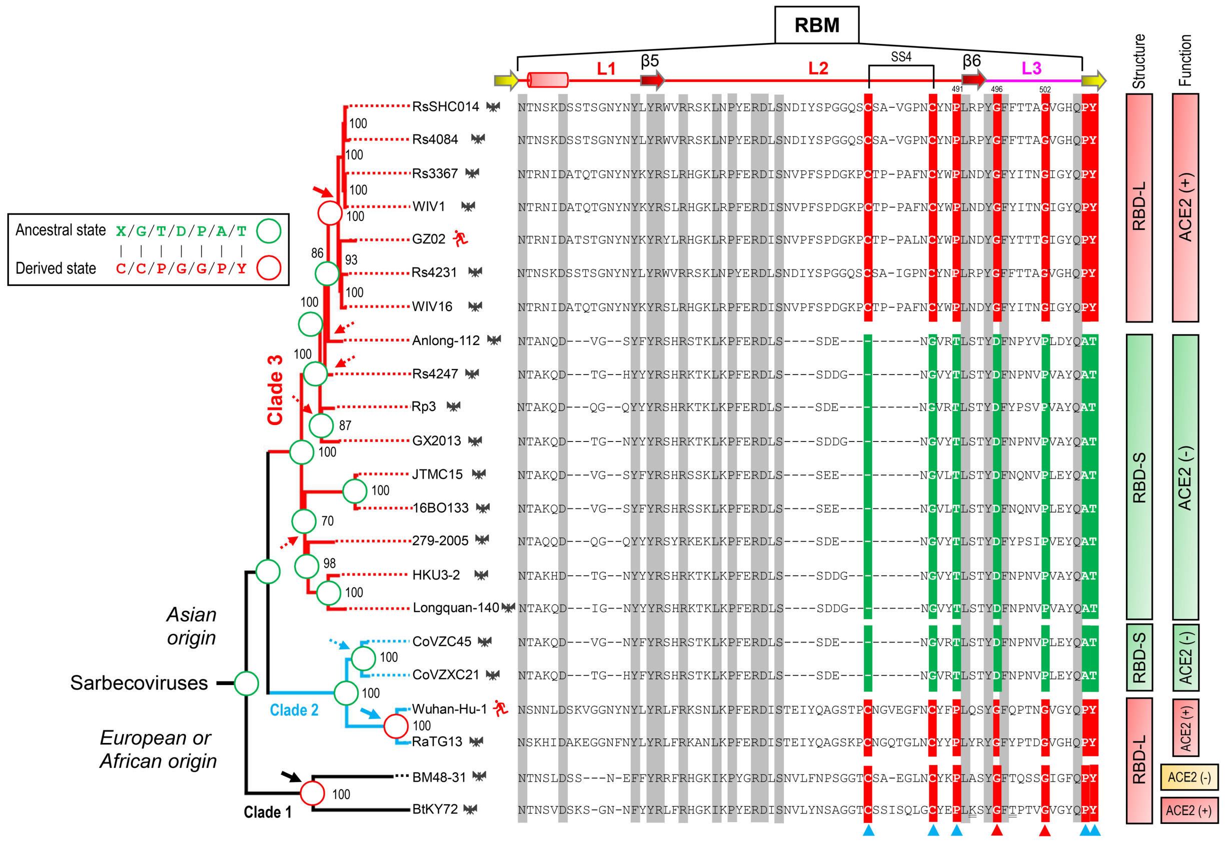


FIGURE 5
Evidence for parallel evolution of ACE2-binding RBMs in three *Sarbecovirus* clades. Phylogenetic relationships of SARS-CoV (strain GZ02), SARS-CoV-2 (Strain Wuhan-Hu-1) and SL-CoVs were reconstructed based on their whole genomes. The relationships were inferred by using the Maximum Likelihood method with the model GTR+G+I. The percentage of trees (≥70%) in which the associated taxa clustered together is shown at the nodes. Proposed insertion or deletion events are denoted by solid or dotted arrows. Right, sequence alignment of RBML1-3 with conservation across the alignment shaded in grey and the determinants for the topology of RBD-based trees in Figure 2B indicated by triangles, in which two directly associated with functional divergence *via* removing unfavorable interactions are shown in red. Sites involved in parallel substitutions are reconstructed for each nodes, in which the same ancestral amino acids are indicated by green circles and the same derived amino acids evolved through independent changes by red circles. "X" denotes a deficient residue. Structure-and function-based classifications are shown on the right of the alignment: ACE2 (+): RBDs capable of binding ACE2; ACE2 (–): RBDs incapable of binding ACE2 (Ren et al., 2008; Hu et al., 2017, 2018; Letko et al., 2020; Roelle et al., 2022; Starr et al., 2022).

be used to explain the histories of the indel mutations in the phylogenetic framework (Figure 5). The first one is three times of independent insertions on three distinct RBD-S-like ancestors which led to the ancestral origins of this trait within sarbecoviruses (Figure 5); the second one is that the common ancestor of sarbecoviruses itself had the insertions and in the subsequent evolution, five times of independent deletions on five distinct RBD-L-like ancestors leading to the loss of the trait (Figure 5). According to the principle of Occam's razor that entities should not be multiplied unnecessarily (Smith, 1980; Orozco-Sevilla and Coselli, 2020) and for a character evolution the fewest changes are the more likely explanation (Futuyma and Kirkpatrick, 2017), we postulated three times of evolutionary insertions other than five times of deletions more likely mediating the origin of ACE2 binding. Moreover, the deletion hypothesis may require a prerequisite, namely, the common ancestor must possess two receptor entry mechanisms because only this can guarantee their survival when the deletion caused the loss of ACE2 binding. By contrast, our insertion hypothesis does not need this prerequisite. In this case, insertion-mediated loop extension provides key functional residues and structural underpinnings for ACE2 binding, as revealed by their functional importance in ACE2 binding (Figures 1, 4). In the phylogenetic framework, seven amino acid sites previously identified as the tree determinants and their mutations related to ACE2 binding can be defined in two different states: an ancestral state described as X/G/T/D/P/A/T ("X" denoting a deficient residue) and a derived state as C/C/P/G/G/P/Y (Figure 5). This is a typical case of parallel substitutions (Storz, 2016), in which independent changes from the ancestral to the derived occurred three times in evolution. In a sub-genome tree reconstructed based on the genomes with their RBD-coding regions deleted (Supplementary Figure 4), these parallel changes were still retained, indicating that this region does not affect the robustness of the genome tree in exploring the evolutionary events. Again, the deletion hypothesis cannot explain why the parallel substitutions observed here still occur after the loss of ACE2 binding although in a reverse manner.

To provide more evidence in support of our hypothesis, we employed an ancestral sequence reconstruction strategy to reconstruct the ancestral states of sarbecovirus evolution with FastML, a method that is especially suitable for the sequences containing indel mutations (see Methods). To minimize the impact of possible recombination, we chose RBM sequences for this end as this region has been predicted to contain no recombination breakpoint (Starr et al., 2022) (Figure 6A). The results show that the ancestral states of the sarbecovirus RBDs (Figure 6B; Supplementary Figure 5) are completely consistent with our hypothesis whatever the genome tree used with or without the RBD-coding region (Figure 6B; Supplementary Figure 4), or different protein substitution models and reconstruction methods used (see Methods). Taken together, our results suggest that the polyphyletic pattern in terms of ACE2 binding in this genome tree is a consequence of ancestral parallel evolution.

## A basal clade-derived RBD incapable of binding hACE2

In the genome tree, BtKY72 and BM48-31 are at the base of the radiation of sarbecoviruses and represent the earliest diverged clade of this group (Figures 5, 6; Supplementary Figure 4). Because they occupy a unique phylogenetic position and their RBDs taxonomically fall into the RBD-L subfamily (Figure 3A), we were interested in studying their potential interaction with ACE2. By using the BtKY72 RBD (abbreviated as BtRBD) as a representative, we prepared its recombinant protein through renaturation from *Escherichia coli* inclusion bodies, which was purified by SEC and identified by HPLC-Q-TOF-MS (Figures 7A,B). Subsequently, we employed SPR, a powerful technique for monitoring the affinity and selectivity of biomolecular interactions, to detect its binding with hACE2. SARS-CoV-2 RBD (Gao and Zhu, 2021) was used as the positive control. In this experiment, hACE2 was covalently linked on the CM5 sensor chip and a RBD protein flowed through the chip surface (Figure 7C). The results showed that the SARS-CoV-2 RBD bound to hACE2 with a $K_D$ of 30.1 nM [association constant ($K_{on}$) of $4.74 \times 10^5$ $M^{-1}$ $s^{-1}$ and dissociation constant ($K_{off}$) of $1.43 \times 10^{-2}$ $s^{-1}$; Figure 7C], which was overall consistent with a previous measurement (Shang et al., 2020). However, BtRBD showed no binding to hACE2 (Figure 7C).

Compared with the ACE2-binding RBDs, BtRBD has two deficient residues in its RBML1 (Supplementary Figure 6). These two residues (Gly-446 and Tyr-449) in the SARS-CoV-2 form three hydrogen bonds with hACE2 (Figures 1B, 8A). Due to the deficiency of these two residues, the BtRBD RBML3 was far away from the interface in its structural model (Figure 8A). This provides a possibility to examine their potential effect on hACE2 binding when introduced into the BtRBD backbone. Using the same strategy described above, we prepared this mutant called BtRBD|GY. Unexpectedly, we found that the insertions of these two residues did not evidently improve the binding of BtRBD to hACE2 (Figure 7C). To provide an explanation of this inability, we evaluated the potential functional importance of these hydrogen bonds to the binding of the SARS-CoV-2 RBD to hACE2 *via* MD simulations. In 100-ns simulations, their survival time was smaller than 10% (Figure 8B), suggesting that they belong to a class of short-lifetime hydrogen bonds. Since the contribution of hydrogen bonds to the stability of proteins is strongly context dependent (Pace et al., 2014), we speculated that these hydrogen bonds could only play a secondary role in ACE2 binding. Alternatively, BtRBD and BtRBD|GY might bind bat ACE2 other than hACE2 given its origin from a bat, as reported recently (Starr et al., 2022). In this case, even minimal binding may be sufficient for viral entry, as observed previously in some bat orthologues of hACE2 that could mediate the infection of SARS-CoV and SARS-CoV-2 (Yan et al., 2021). Since the binding and susceptibility are not always consistent, and the ability to support the entry of virus is much more important than the binding in terms of susceptibility to virus infection, more studies to address the significance of the two-residue insertion in BtKY72 infection are needed in the future.

# Discussion

In this work, we have provided multidimensional evidence in support of the role of parallel insertions-and point mutations-driven functional innovation in the ancestral origins of ACE2-utilizing sarbecoviruss. Parallel evolution occurring in multiple evolutionary lineages of viruses are not uncommon (Gutierrez et al., 2019), especially those that register frequent cross-species transmission events (Longdon et al., 2014; Gutierrez et al., 2019). A recent study also showed that the emergence of highly pathogenic avian influenza A viruses is a result of parallel evolution (Escalera-Zamudio et al., 2020). Multiple mechanisms have been proposed to explain such parallel evolution in viruses, such as point mutations involved in the development of antiviral drug resistance, adaptation to new host species, and evasion of host immunity (Gutierrez et al., 2019). For example, a Glu to Lys change at position 627 of PB2 increased virulence on mammalian hosts, in both H5N1 and H3N2 subtypes (Steel et al., 2009). In addition to viruses, parallel evolution has also been documented in animals. For instance, parallel amino acid replacements have resulted in acquired enhanced digestive efficiencies in Asian and African leaf-eating monkeys (Prud'homme and Carroll, 2006; Zhang, 2006). The independent development of closely corresponding adaptive features in two or more groups of mammals that occupy different but equivalent habitats has also been reported previously (Storz, 2016).

Based on the phylogenetic conflict between two trees built from different gene segments, it has been proposed that recombination-mediated exchange of spike RBDs plays a role in the CoV evolution (Boni et al., 2020; Wells et al., 2021). But as mentioned in Introduction, this opinion remains controversial especially in the explanation of the origin of SARS-CoV-2 (Boni et al., 2020; Xiao et al., 2020) and such recombination could not explain how the first ACE2-utilizing sarbecoviruses originated because the RBM directly involved in interaction with ACE2 is not a mosaic organization produced by recombination, as evidenced by the lack of a recombination breakpoint in this region (Starr et al., 2022). We found that for the phylogenetic conflict between the RBD tree and the genome tree, it is more likely explained by parallel evolution-mediated functional clustering of the RBD-L proteins in the RBD tree (Figure 3), which can be recognized by analysis of amino acid changes in the framework of a genome tree and further strengthened by ancestral sequence reconstruction. This well explains the origins of the first ACE2-utilizing sarbecoviruses. The parallel events repeatedly occurred in the evolution of the SARS-CoV and SARS-CoV-2 clades included: (1) Insertion-mediated loop extensions in RBML1 and RBML2. Such extensions created new structural basis through contribution of key structural and functional residues involved in interactions with ACE2 and assembly of one new

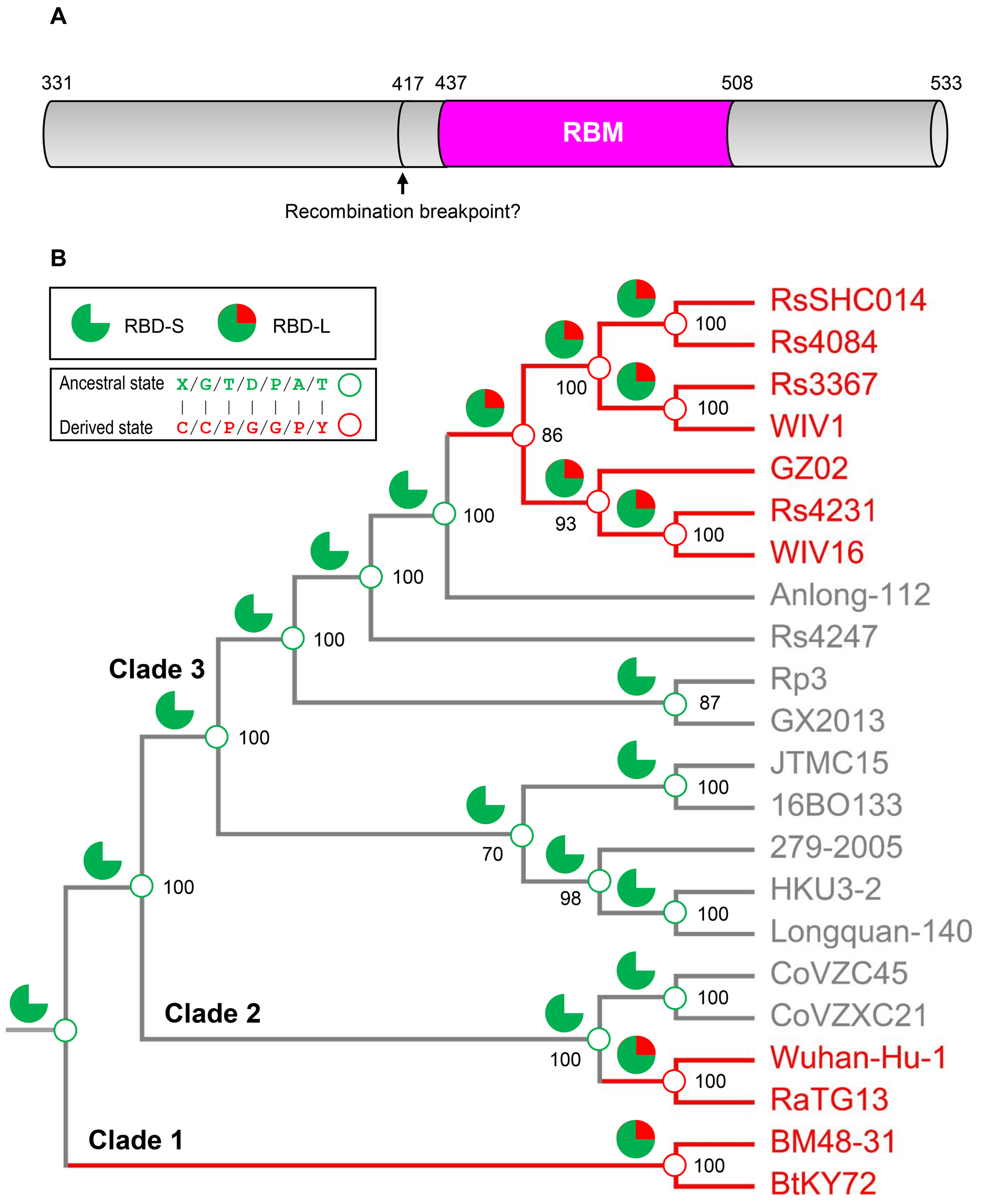


FIGURE 6
Ancestral sequence reconstruction for elucidating the histories of the insertions and point mutations proposed by parallel evolution. **(A)** Schematic diagram illustrating the structural organization of RBDs, in which the pink region corresponding to the RBM was used to reconstruct ancestral sequences. The putative recombination breakpoint was denoted by an arrow (Starr et al., 2022). Residues were numbered according to the SARS-CoV-2 spike. **(B)** The projection of the ancestral traits including the insertions and point mutations proposed by parallel evolution onto the genome phylogeny. Green sector graphs represent RBD-S without an insertion and green and red sector graphs represent RBD-L with the insertion. Green and red circles represent an ancestral state and a derived state, respectively.

disulfide bridge modulating the dynamics of RBML2; (2) Insertion-driven substitution rate increase in RBML3 (Figure 3C). These mutations remove energetically unfavorable interactions with ACE2 and affect the dynamics and conformations of the key functional RBML2 (Figure 4). Our observations suggest a role of correlated evolution among different loops in the emergence of ACE2-utilizing sarbecoviruses. Modifications of ancestral loops by molecular tinkering are also in line with the opinion that loops in an ancestral structure are targets for indel mutations during evolution (Pascarella and Argos, 1992).

Although the events all also occurred in clade 1 (Figures 5, 6), some of its members could not bind ACE2 (Letko et al., 2020; Roelle et al., 2022; Starr et al., 2022). This is likely due to several residues deficiency in the two loops (RBML1 and RBML2), as identified by their length falling between the long and short RBDs. However, adding the deficient residues, as in the case of BtRBD|GY, did not improve the ACE2 binding

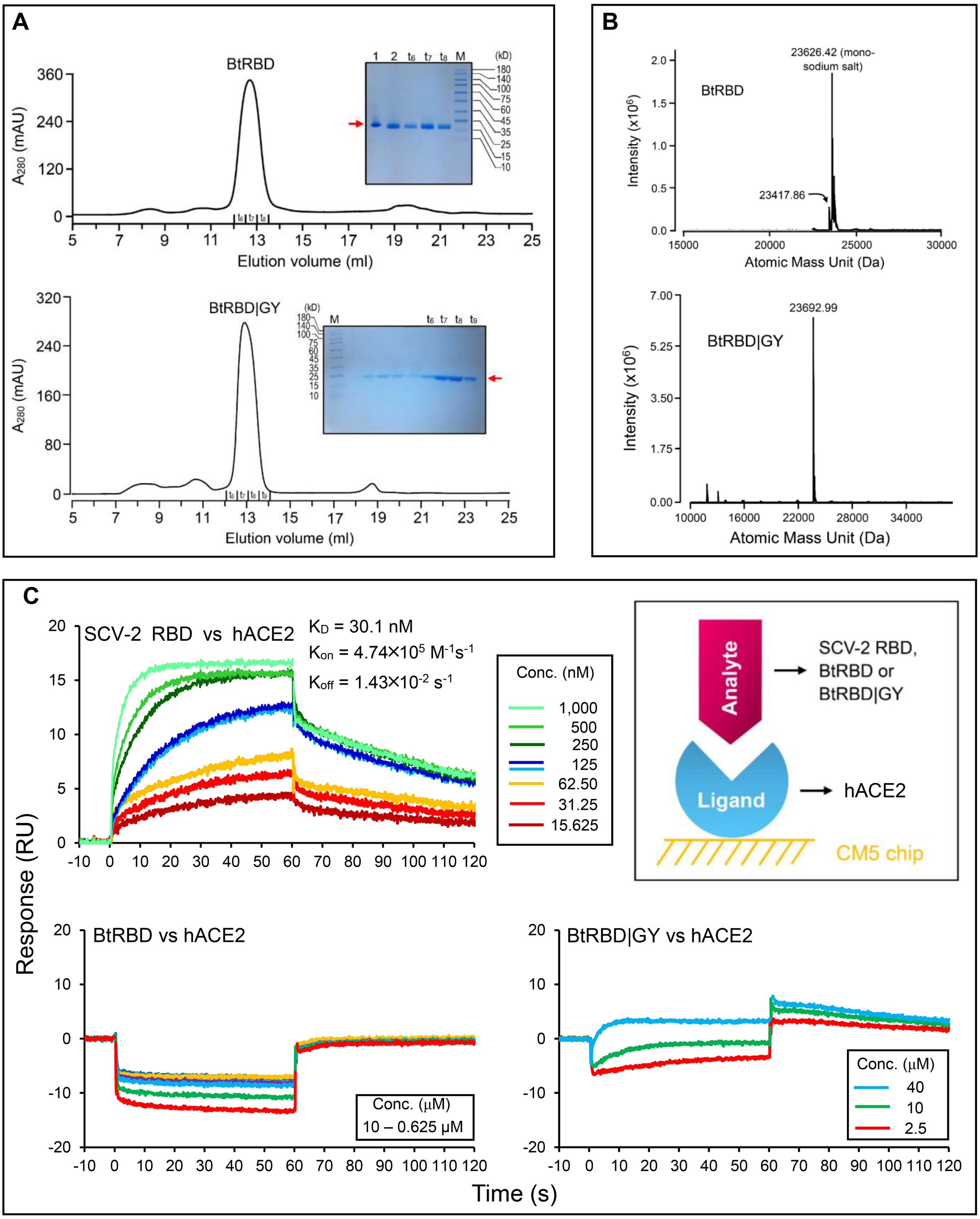


FIGURE 7
Purification, identification and functional characterization of recombinant RBDs. **(A)** Purification of refolded BtRBD and BtRBD|GY by SEC. Inset: SDS-PAGE analysis of the purified products, marked by a red arrow. "t$_6$ to t$_9$" denote collection tubes in SEC and "M" denotes protein molecular weight standard. **(B)** HPLC-Q-TOF-MS determining the molecular mass of BtRBD and BtRBD|GY. **(C)** Sensorgrams of SARS-CoV-2 RBD binding to the ACE2-immobilized chip surface (left top). The 125 nM analyte concentration was analyzed in duplicate. The concentrations used were 1,000–15.625nM with two-fold serial dilutions. Sensorgrams of BtRBD to the chip surface (left bottom). The concentrations used were 10,000–625 nM with two-fold serial dilutions. Sensorgrams of BtRBD|GY to the chip surface (right bottom). The concentrations used were 40,000, 10,000, and 2,500nM. Inset, schematic diagram of SPR experiment, in which the ligand hACE2 was covalently immobilized onto CM5 *via* its amine groups and the analytes (RBDs) flowed over the surface.

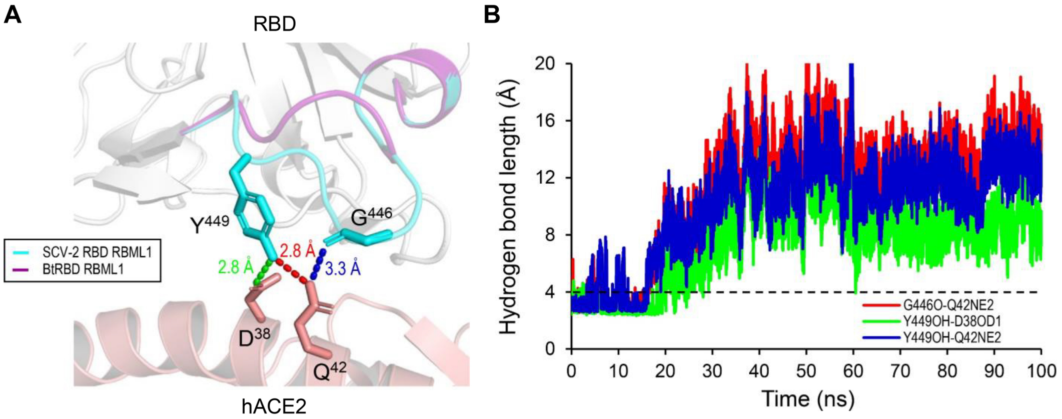


FIGURE 8
The structural basis of the RBML1 of SARS-CoV-2 RBD interacting with hACE2. **(A)** Gly-446 and Tyr-449 of the RBML1 (colored in cyan) interact with Gln-42 and Asp-38 of hACE2 *via* three hydrogen bonds (pdb entry 6LZG). In the model of BtRBD, its RBML1 far away from the interface is colored in purple. **(B)** The lifetimes of the hydrogen bonds during 100-ns MD simulations. The dashed line represents the length threshold (4 Å) of a hydrogen bond.

of BtRBD. It has been found that the development of ACE2 binding on the scaffolds of BM48-31 and short RBDs requires replacing all 14 contact points and the surrounding amino acids in the RBM (Letko et al., 2020). This highlights the role of non-interacting residues in ACE2 binding. During submission of this manuscript, we noticed two recent publications that reported the binding function of BtKY72 RBD to human and bat ACE2 (Roelle et al., 2022; Starr et al., 2022). Our finding that this RBD was unable to bind hACE2 is consistent with (Starr et al., 2022) but different from (Roelle et al., 2022) that recorded some activity on hACE2. Such discrepancy could be due to the difference in the assay methods used (SPR vs. mixed cell pseudotyped virus infection assay; Roelle et al., 2022). Interestingly, this RBD can bind two bat-derived RBDs (Starr et al., 2022). Collectively, these observations suggest that the clade 1 CoVs might represent an evolutionary intermediate linking ACE2 utilizing and non-utilizing sarbecoviruses. We thus propose that parallel evolution in sarbecoviruses could involve a state of evolutionary intermediates (Figure 9). The parallel fixation of key amino acids in these intermediates with different genetic backgrounds might be the first step in an adaptive walk (Storz, 2016) *via* exerting a favorable effect on the mutational pathways of spike protein evolution into ACE2 binding by sequence optimization, as seen in BtRBD whose mutations (K493Y and T498W) enabled it to interact with hACE2 (Starr et al., 2022). If this is true, it suggests that the emergence of ACE2 binding has evolved gradually and repeatedly through molecular tinkering of a pre-existing progenitor over an extended period, as the proposed case for the evolution of the antibody-based immune system (Klein and Nikolaidis, 2005). This suggestion is also highly consistent with the opinion that evolution is often gradual (Futuyma and Kirkpatrick, 2017).

The emergence of a trait from an evolutionary point of view is unlikely to originate more than once by chance and therefore three times of independent origins of ACE2 binding must have been driven by a common selective pressure. Although it is known that viruses and their hosts are locked in an evolutionary arm race (Yap et al., 2020), the fact the ancestral sarbecoviruses still infected bats after they had evolutionarily gained ACE2 binding suggests that the development of the trait is more likely to commonly deal with the insertion-caused decrease in the binding ability of their RBDs to the unknown host receptor other than to circumvent the bats' defences due to resistance acquirement by the hosts in the arm race. This can be considered as a compensation mechanism during virus evolution and represents an example of mutation-driven evolution of new function (Nei, 2013). A new study provides further support for this opinion. In this study, it was found that the evolutionary gain of an insert in the loop of a nematode defensin leads to the emergence of a new antibacterial function (Gu et al., 2022). Such an insertion event also independently occurred in its ortholog from a genetically distant nematode species (Gu et al., 2018). In particular, our opinion can overall satisfy all four criteria regarding parallel adaptive evolution at the protein sequence level (Zhang, 2006): (1) Similar changes in RBD function occur in three independent evolutionary clades; (2) Parallel amino acid mutations (both insertion and substitution) are observed in these RBDs; (3) A compensation mechanism in receptor usage likely commonly driving their evolution; (4) The parallel mutations are responsible for the parallel emergence of ACE2 binding.

It is worth mentioning that our finding that distantly related coronaviruses independently evolve ACE2 binding in their respective ancestors *via* insertions to increase the flexibility of the functional loop involved in interaction with ACE2, and point mutations to remove unfavorable interactions between RBD and ACE2 is very similar to the evolution of certain toxins. One well-documented example is that insectivorous mammals and lizards both independently evolved their toxins from a class of homologous, ancestral non-toxic enzymes by insertions to increase the flexibility of functional loops and point mutations to introduce new chemical environment (Aminetzach et al., 2009). Also, loop extension and key point mutations were found to jointly drive the emergence of scorpion sodium channel toxins from an ancestral defensin scaffold (Zhu et al., 2020). Although there is no comparability between viral spike proteins and animal toxins, they both may have evolved to use a common strategy to make their weapons.

Different from SARS-CoV and SARS-CoV-2 that both gained receptor binding by parallel evolution to target ACE2, another human coronavirus - MERS-CoV is known to utilize dipeptidyl peptidase 4 (DPP4) instead of ACE2 as the host receptor, which involves the S1 CTD of the spike protein as RBD (Millet et al., 2021). Although these three CoVs all belong to β-coronaviruses and infect humans, the evolutionary mechanisms of their receptor binding origins are different.

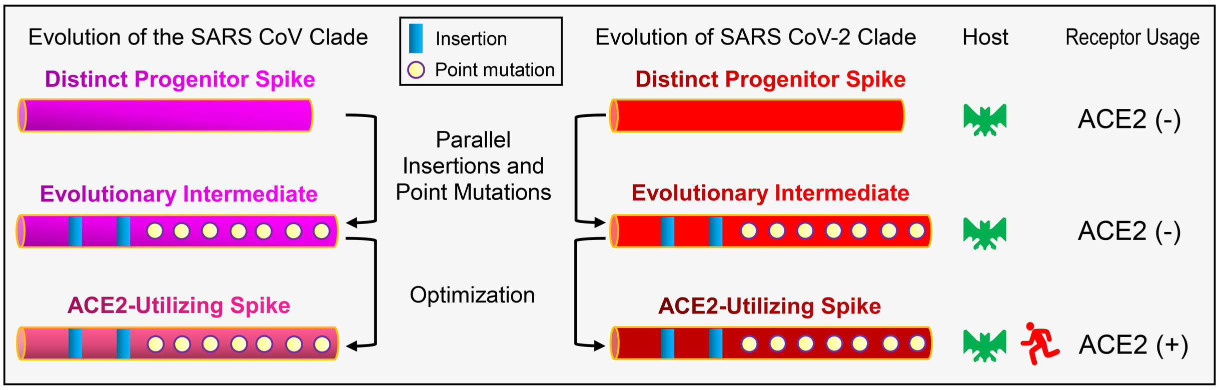

Evolution of the SARS CoV Clade
Insertion
Point mutation
Evolution of SARS CoV-2 Clade
Host
Receptor Usage
Distinct Progenitor Spike
Distinct Progenitor Spike
Parallel Insertions and Point Mutations
ACE2 (-)
Evolutionary Intermediate
Evolutionary Intermediate
ACE2 (-)
Optimization
ACE2-Utilizing Spike
ACE2-Utilizing Spike
ACE2 (+)

FIGURE 9
Schematic diagram of the proposed evolutionary histories of ACE2-utilizing spike proteins in sarbecoviruses. For simplicity, only RBDs are shown here. In this evolutionary scenario, two distinct progenitors are developed into SARS-CoV and SARS-CoV-2 clades *via* parallel insertions and point mutations followed by sequence optimization. ACE2 utilizing and non-utilizing are denoted by minus and plus signs, respectively.

For MERS-CoV, its spike RBD involved in DPP4 binding (Wang et al., 2013; Xu et al., 2020) exhibits a rather low sequence similarity to the RBDs of other two CoVs involved in ACE2 binding. This could be a consequence of divergent evolution after speciation, which occurred from a common ancestor *via* point mutations and an insertion mutation (Wang et al., 2013; Xu et al., 2020) to target a different host receptor.

Finally, our work highlights the importance of an integrative approach utilizing multidimensional data in exploring the molecular origins of specific phenotypes of viruses from their genotypes. Given that ACE2 is also convergently targeted by HCoV-NL63, a human α-CoV with a similar but distinct ACE2 binding mode from that of β-CoVs (Rawat et al., 2021), our approach is likely to be useful in studying how it originates within the α-CoVs.

## 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

SZ conceived and designed this study and performed evolutionary analysis and molecular dynamics simulations. BG performed experiments. BG and SZ commonly wrote the paper. All authors contributed to the article and approved the submitted version.

## 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.2023.1118025/full#supplementary-material

## References

Aminetzach, Y. T., Srouji, J. R., Kong, C. Y., and Hoekstra, H. E. (2009). Convergent evolution of novel protein function in shrew and lizard venom. *Curr. Biol.* 19, 1925–1931. doi: 10.1016/j.cub.2009.09.022

Arenas, M., Araujo, N. M., Branco, C., Castelhano, N., Castro-Nallar, E., and Pérez-Losada, M. (2018). Mutation and recombination in pathogen evolution: relevance, methods and controversies. *Infect. Genet. Evol.* 63, 295–306. doi: 10.1016/j.meegid.2017.09.029

Arya, R., Kumari, S., Pandey, B., Mistry, H., Bihani, S. C., Das, A., et al. (2021). Structural insights into SARS-CoV-2 proteins. *J. Mol. Biol.* 433:166725:166725. doi: 10.1016/j.jmb.2020.11.024

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., et al. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40, W580–W584. doi: 10.1093/nar/gks498

Bolles, M., Donaldson, E., and Baric, R. S. (2011). SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission. *Curr. Opin. Virol.* 1, 624–634. doi: 10.1016/j.coviro.2011.10.012

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* 5, 1408–1417. doi: 10.1038/s41564-020-0771-4

Brook, C. E., and Dobson, A. P. (2015). Bats as 'special' reservoirs for emerging zoonotic pathogens. *Trends Microbiol.* 23, 172–180. doi: 10.1016/j.tim.2014.12.004

Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9

Dobson, A. P. (2005). What links bats to emerging infectious diseases? *Science* 310, 628–629. doi: 10.1126/science.1120872

Drexler, J. F., Gloza-Rausch, F., Glende, J., Corman, V. M., Muth, D., Goettsche, M., et al. (2010). Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *J. Virol.* 84, 11336–11349. doi: 10.1128/JVI.00650-10

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8, 163–167. doi: 10.1101/gr.8.3.163

Escalera-Zamudio, M., Golden, M., Gutiérrez, B., Thézé, J., Keown, J. R., Carrique, L., et al. (2020). Parallel evolution in the emergence of highly pathogenic avian influenza a viruses. *Nat. Commun.* 11:5511:5511. doi: 10.1038/s41467-020-19364-x

Futuyma, D. J., and Kirkpatrick, M. (2017). *Evolution. 4th* Edn Sinauer Associates Sinauer Associates, Inc., USA, Massachusetts.

Gao, B., and Zhu, S. (2021). A fungal defensin targets the SARS-CoV-2 spike receptor-binding domain. *J. Fungi* 7:553. doi: 10.3390/jof7070553

Ge, X. Y., Li, J. L., Yang, X. L., Chmura, A. A., Zhu, G., Epstein, J. H., et al. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* 503, 535–538. doi: 10.1038/nature12711

Gu, J., Gao, B., and Zhu, S. (2018). Characterization of bi-domain drosomycin-type antifungal peptides in nematodes: an example of convergent evolution. *Dev. Comp. Immunol.* 87, 90–97. doi: 10.1016/j.dci.2018.06.005

Gu, J., Isozumi, N., Gao, B., Ohki, S., and Zhu, S. (2022). Mutation-driven evolution of antibacterial function in an ancestral antifungal scaffold: significance for peptide engineering. *Front. Microbiol.* 13:1053078. doi: 10.3389/fmicb.2022.1053078

Gutierrez, B., Escalera-Zamudio, M., and Pybus, O. G. (2019). Parallel molecular evolution and adaptation in viruses. *Curr. Opin. Virol.* 34, 90–96. doi: 10.1016/j.coviro.2018.12.006

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* 19, 409–424. doi: 10.1038/s41579-021-00573-0

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cells* 181, 271–280.e8. doi: 10.1016/j.cell.2020.02.052

Hogg, P. J. (2003). Disulfide bonds as switches for protein function. *Trends Biochem. Sci.* 28, 210–214. doi: 10.1016/S0968-0004(03)00057-4

Hu, B., Zeng, L. P., Yang, X. L., Ge, X. Y., Zhang, W., Li, B., et al. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* 13:e1006698. doi: 10.1371/journal.ppat.1006698

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., et al. (2018). Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.* 7:154, 1–10. doi: 10.1038/s41426-018-0155-5

Hulswit, R. J. G., de Haan, C. A. M., and Bosch, B.-J. (2016). Coronavirus spike protein and tropism changes. *Adv. Virus Res.* 96, 29–57. doi: 10.1016/bs.aivir.2016.08.004

Klein, J., and Nikolaidis, N. (2005). The descent of the antibody-based immune system by gradual evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102, 169–174. doi: 10.1073/pnas.0408480102

Kuhn, J. H., Li, W., Choe, H., and Farzan, M. (2004). Angiotensin-converting enzyme 2: a functional receptor for SARS coronavirus. *Cell. Mol. Life Sci.* 61, 2738–2743. doi: 10.1007/s00018-004-4242-5

Laskowski, R. A., and Swindells, M. B. (2011). LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* 51, 2778–2786. doi: 10.1021/ci200227u

Letko, M., Marzi, A., and Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* 5, 562–569. doi: 10.1038/s41564-020-0688-y

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annu. Rev. Virol.* 3, 237–261. doi: 10.1146/annurev-virology-110615-042301

Li, P., Guo, R., Liu, Y., Zhang, Y., Hu, J., Ou, X., et al. (2021). The Rhinolophus affinis bat ACE2 and multiple animal orthologs are functional receptors for bat coronavirus RaTG13 and SARS-CoV-2. *Sci. Bull.* 66, 1215–1227. doi: 10.1016/j.scib.2021.01.011

Li, F., Li, W., Farzan, M., and Harrison, S. C. (2005). Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* 309, 1864–1868. doi: 10.1126/science.1116480

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., et al. (2005a). Bats are natural reservoirs of SARS-like coronaviruses. *Science* 310, 676–679. doi: 10.1126/science.1118391

Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., et al. (2005b). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* 24, 1634–1643. doi: 10.1038/sj.emboj.7600640

Liu, K., Pan, X., Li, L., Yu, F., Zheng, A., Du, P., et al. (2021). Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. *Cells* 184, 3438–3451.e10. doi: 10.1016/j.cell.2021.05.031

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., and Jiggins, F. M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog.* 10:e1004395. doi: 10.1371/journal.ppat.1004395

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, 565–574. doi: 10.1016/S0140-6736(20)30251-8

McDonald, I. K., and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238, 777–793. doi: 10.1006/jmbi.1994.1334

Millet, J. K., Jaimes, J. A., and Whittaker, G. R. (2021). Molecular diversity of coronavirus host cell entry receptors. *FEMS Microbiol. Rev.* 45:fuaa057. doi: 10.1093/femsre/fuaa057

Nakagawa, K., Lokugamage, K. G., and Makino, S. (2016). Viral and cellular mRNA translation in coronavirus-infected cells. *Adv. Virus Res.* 96, 165–192. doi: 10.1016/bs.aivir.2016.08.001

Nei, M. (2013) *Mutation-driven evolution*. Oxford University Press, UK, Oxford

Nikolovska-Coleska, Z. (2015). Studying protein-protein interactions using surface plasmon resonance. *Methods Mol. Biol.* 1278, 109–138. doi: 10.1007/978-1-4939-2425-7_7

Orozco-Sevilla, V., and Coselli, J. S. (2022). Commentary: Occam's razor: the simplest solution is always the best. *J. Thorac. Cardiovasc. Surg.* 164, 1053–1054. doi: 10.1016/j.jtcvs.2020.10.087

Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Schell, D., et al. (2014). Contribution of hydrogen bonds to protein stability. *Protein Sci.* 23, 652–661. doi: 10.1002/pro.2449

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* 224, 461–471. doi: 10.1016/0022-2836(92)91008-d

Piplani, S., Singh, P. K., Winkler, D. A., and Petrovsky, N. (2021). In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Sci. Rep.* 11:13063. doi: 10.1038/s41598-021-92388-5

Prud'homme, B., and Carroll, S. B. (2006). Monkey see, monkey do. *Nat. Genet.* 38, 740–741. doi: 10.1038/ng0706-740

Rawat, P., Jemimah, S., Ponnuswamy, P. K., and Gromiha, M. M. (2021). Why are ACE2 binding coronavirus strains SARS-CoV/SARS-CoV-2 wild and NL63 mild? *Proteins* 89, 389–398. doi: 10.1002/prot.26024

Ren, W., Qu, X., Li, W., Han, Z., Yu, M., Zhou, P., et al. (2008). Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin. *J. Virol.* 82, 1899–1907. doi: 10.1128/JVI.01085-07

Roelle, S. M., Shukla, N., Pham, A. T., Bruchez, A. M., and Matreyek, K. A. (2022). Expanded ACE2 dependencies of diverse SARS-like coronavirus receptor binding domains. *PLoS Biol.* 20:e3001738. doi: 10.1371/journal.pbio.3001738

Schmidpeter, P. A., and Schmid, F. X. (2015). Prolyl isomerization and its catalysis in protein folding and protein function. *J. Mol. Biol.* 427, 1609–1631. doi: 10.1016/j.jmb.2015.01.023

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., et al. (2020). Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221–224. doi: 10.1038/s41586-020-2179-y

Shi, Z., and Wang, L. F. (2011). Evolution of SARS coronavirus and the relevance of modern molecular epidemiology. *Genet. Evol. Infect. Dis.* 2017, 601–619. doi: 10.1016/B978-0-12-799942-5.00026-3

Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.* 92, 452–456. doi: 10.1073/pnas.92.2.452

Siu, Y. L., Teoh, K. T., Lo, J., Chan, C. M., Kien, F., Escriou, N., et al. (2008). The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J. Virol.* 82, 11318–11330. doi: 10.1128/JVI.01052-08

Smith, T. F. (1980). Occam's razor. *Nature* 285:620. doi: 10.1038/285620a0

Starr, T. N., Zepeda, S. K., Walls, A. C., Greaney, A. J., Alkhovsky, S., Veesler, D., et al. (2022). ACE2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature* 603, 913–918. doi: 10.1038/s41586-022-04464-z

Steel, J., Lowen, A. C., Mubareka, S., and Palese, P. (2009). Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Pathog.* 5:e1000252. doi: 10.1371/journal.ppat.1000252

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* 17, 239–250. doi: 10.1038/nrg.2016.11

Tao, Y., and Tong, S. (2019). Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. *Microbiol. Resour. Announc.* 8, e00548–e00519. doi: 10.1128/MRA.00548-19

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455, 105–108. doi: 10.1038/nature07175

Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D. S. (2008). How protein stability and new functions trade off. *PLoS Comput. Biol.* 4:e1000002. doi: 10.1371/journal.pcbi.1000002

Valencia, A., and Pazos, F. (2003). Prediction of protein-protein interactions from evolutionary information. *Methods Biochem. Anal.* 44, 411–426. doi: 10.1002/0471721204.CH20

Verano-Braga, T., Martins, A. L. V., Motta-Santos, D., Campagnole-Santos, M. J., and Santos, R. A. S. (2020). ACE2 in the renin-angiotensin system. *Clin. Sci. (Lond.)* 134, 3063–3078. doi: 10.1042/CS20200478

Wang, N., Shi, X., Jiang, L., Zhang, S., Wang, D., Tong, P., et al. (2013). Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell Res.* 23, 986–993. doi: 10.1038/cr.2013.92

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., et al. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cells* 181, 894–904.e9. doi: 10.1016/j.cell.2020.03.045

Wells, H. L., Letko, M., Lasso, G., Ssebide, B., Nziza, J., Byarugaba, D. K., et al. (2021). The evolutionary history of ACE2 usage within the coronavirus subgenus Sarbecovirus. *Virus Evol.* 7: veab007. doi: 10.1093/ve/veab007

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J. J., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* 583, 286–289. doi: 10.1038/s41586-020-2313-x

Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., et al. (2020). Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* 63, 457–460. doi: 10.1007/s11427-020-1637-5

Yan, H., Jiao, H., Liu, Q., Zhang, Z., Xiong, Q., Wang, B. J., et al. (2021). ACE2 receptor usage reveals variation in susceptibility to SARS-CoV and SARS-CoV-2 infection among bat species. *Nat. Ecol. Evol.* 5, 600–608. doi: 10.1038/s41559-021-01407-1

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367, 1444–1448. doi: 10.1126/science.abb2762

Yap, M. W., Young, G. R., Varnaite, R., Morand, S., and Stoye, J. P. (2020). Duplication and divergence of the retrovirus restriction gene Fv1 in Mus caroli allows protection from multiple retroviruses. *PLoS Genet.* 16:e1008471. doi: 10.1371/journal.pgen.1008471

Yoshida, R., and Nei, M. (2016). Efficiencies of the NJp, maximum likelihood, and bayesian methods of phylogenetic construction for compositional and noncompositional genes. *Mol. Biol. Evol.* 33, 1618–1624. doi: 10.1093/molbev/msw042

Zaman, S., Sledzieski, S., Berger, B., Wu, Y., and Bansal, M. S. (2021). Phylogenetic reconciliation reveals extensive ancestral recombination in Sarbecoviruses and the SARS-CoV-2 lineage. *bioRxiv* 2021.08.12.456131

Zhang, J. (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat. Genet.* 38, 819–823. doi: 10.1038/ng1812

Zhang, Z., Huang, J., Wang, Z., Wang, L., and Gao, P. (2011). Impact of indels on the flanking regions in structural domains. *Mol. Biol. Evol.* 28, 291–301. doi: 10.1093/molbev/msq196

Zhu, S., Gao, B., Peigneur, S., and Tytgat, J. (2020). How a scorpion toxin selectively captures a prey sodium channel: the molecular and evolutionary basis uncovered. *Mol. Biol. Evol.* 37, 3149–3164. doi: 10.1093/molbev/msaa152

Zhu, S., Gao, B., Umetsu, Y., Peigneur, S., Li, P., Ohki, S., et al. (2022). Adaptively evolved human oral actinomyces-sourced defensins show therapeutic potential. *EMBO Mol. Med.* 14:e14499. doi: 10.15252/emmm.202114499

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fmicb.2023.1147652

OPEN ACCESS

EDITED BY
Jianying Gu,
College of Staten Island,
United States

REVIEWED BY
A. S. M. Rubayet Ul Alam,
Jashore University of Science and Technology,
Bangladesh
Jia Weixin,
South China Agricultural University,
China

*CORRESPONDENCE
Yuanyuan Zhu
✉ zhuyuanyzz@163.com

SPECIALTY SECTION
This article was submitted to
Virology,
a section of the journal
Frontiers in Microbiology

RECEIVED 19 January 2023
ACCEPTED 16 February 2023
PUBLISHED 09 March 2023

CITATION
Li F, Li Y, Ma J, Wu R, Zou X, Liu Y, Zhao Q and Zhu Y (2023) Molecular evolution, diversity, and adaptation of foot-and-mouth disease virus serotype O in Asia.
*Front. Microbiol.* 14:1147652.
doi: 10.3389/fmicb.2023.1147652

# Molecular evolution, diversity, and adaptation of foot-and-mouth disease virus serotype O in Asia

Fangtao Li, Yan Li, Jianrong Ma, Ruizhi Wu, Xingqi Zou, Yebing Liu, Qizu Zhao and Yuanyuan Zhu*

National/WOAH Reference Laboratory for Classical Swine Fever, China Institute of Veterinary Drug Control, Beijing, China

Foot-and-mouth disease (FMD) is highly contagious and affects the economy of many countries worldwide. Serotype O is the most prevalent and is present in many regions of Asia. Lineages O/SEA/Mya-98, O/Middle East-South Asia (ME-SA)/PanAsia, O/Cathay and O/ME-SA/Ind-2001 have been circulating in Asian countries. Low antigenic matching between O/Cathay strains and current vaccine strains makes the disease difficult to control, therefore, analyzing the molecular evolution, diversity, and host tropisms of FMDV Serotype O in Asia may be helpful. Our results indicate that Cathay, ME-SA, and SEA are the predominant topotypes of FMDV serotype O circulating in Asia in recent years. Cathay topotype FMDV evolves at a higher rate compared with ME-SA and SEA topotypes. From 2011 onwards, the genetic diversity of the Cathay topotype has increased substantially, while large reductions were found in the genetic diversity of both ME-SA and SEA topotypes, suggesting a trend that infections sustained by the Cathay topotype were becoming a more severe epidemic in recent years. Analyzing the distributions of host species through time in the dataset, we found that the O/Cathay topotype was characterized by a highly swine-adapted tropism in contrast with a distinct host preference for O/ME-SA. The O/SEA topotype strains identified in Asia were isolated mainly from cattle until 2010. It is worth noting that there may be a fine-tuned tropism of the SEA topotype viruses for host species. To further explore the potential molecular mechanism of host tropism divergence, we analyzed the distribution of structure variations on the whole genome. Our findings suggest that deletions in the PK region may reflect a common pattern of altering the host range of serotype O FMDVs. In addition, the divergence of host tropism may be due to accumulated structural variations across the viral genome, rather than a single indel mutation.

KEYWORDS

**FMDV, O/Cathay, genetic diversity, host tropisms, potential molecular mechanism**

## 1. Introduction

FMD is a contagious viral disease clinically characterized by lesions in the mouth and feet of cloven-hoofed animals, which has affected more than 70 animal species including cattle, pigs, sheep, goats, water buffalo, and wild ruminants (Garcia, 2012; Li et al., 2021), and imposes burden on the economy of many countries worldwide (Sinkala et al., 2014; Diaz-San Segundo et al., 2017). FMD virus (FMDV) is a member of the Aphthovirus genus within the Picornaviridae family. It is ~30 nm in diameter and forms an icosahedral structure with a sedimentation coefficient of 146S, and consists of 60 copies of each of the capsid proteins VP1–VP4 (Knowles and Samuel, 2003). The capsid protein precursor (P1 + 2A) of picornavirus is initially digested into VP0, VP1, and VP3 to form a protomer. Mature virions are ultimately formed with the package of the genome into capsids and the autocatalytic cleavage of VP0 into VP2 and VP4 (Freimanis et al., 2016).

FMDV has seven antigenically distinct serotypes, namely, O, A, C, Asia 1, Southern African Territories (SAT) 1, SAT 2, SAT 3 and numerous subtypes (Doel, 2003). FMDV serotype O is one of the global epidemic serotypes and causes significant economic loss (Shao et al., 2011). Three serotypes of FMDV, including serotypes O, A, and Asia 1, have caused epidemics in Asia, making FMD difficult to control (Brito et al., 2017; Blacksell et al., 2019). The serotype O is the most prevalent of the three serotypes and is present in many regions in Asia (Zhu et al., 2019). In India, FMDV serotype O dominated the outbreak scenario, accounting for about 92% of all outbreaks (Subramaniam et al., 2022). Serotypes O is also the most prevalent in Bangladesh, and the maximum outbreaks occurred with this serotype (Hossen et al., 2020). Despite the efforts of the National FMD Control Program, A large majority of the reported outbreaks caused by FMDV serotype O in Nepal remains a major threat to the livestock industry in Nepal (Adhikari et al., 2018). Along with the three usual strains O/SEA/Mya-98, O/Middle East-South Asia (ME-SA)/PanAsia, and O/Cathay, an emerging O/ME-SA/Ind-2001 (ME-SA) lineage has been circulating in this region since 2015 and have now spread to most of the Southeast Asian countries (Upadhyaya et al., 2021). In recent years, serotype O FMDVs have been mainly responsible for outbreaks of FMD in China (Li et al., 2022).

In 1997, a FMDV confirmed in Taiwan showed atypical pathogenicity with high morbidity and mortality in swine but no effect on cattle, leading to severe economic losses (Dunn and Donaldson, 1997). The causative agent was confirmed to be a distinct topotype of serotype O (i.e., O/Cathay), which was identified for the first time in 1970 in China (Beard and Mason, 2000). Since the catastrophic outbreak in Taiwan, sporadic outbreaks caused by O/Cathay strains have been reported in China and several Southeast Asian countries, together with O/SEA/Mya-98, O/ME-SA/PanAsia, and O/ME-SA/Ind-2001 strains (Brito et al., 2017).

Although vaccination is the key to control serotype O FMD, the available vaccines are not able to provide enough cross-protection as outbreaks still occurred despite repeated vaccinations (Mahapatra et al., 2017; Lee et al., 2020; Park et al., 2021). The vaccines used showed a good match with the O/SEA and O/ME-SA viruses, whereas none of the recently circulating O/Cathay viruses were protected by any of the vaccine strains, including the existing O/Cathay vaccine, indicating an antigenic drift and the urgency to develop new vaccine strains (Upadhyaya et al., 2021). Low antigenic matching between the O/Cathay strains and current vaccine strain makes the disease difficult to control, so current strategies to eradicate FMDV of this topotype rely on the rapid detection of infected animals and control measures including movement restriction and culling of animals suspected of infection (Nishi et al., 2021). Therefore, it is necessary to analyze the molecular evolution and host tropisms of FMDV serotype O in Asia.

This study investigated the molecular epidemiology, evolutionary dynamics, and host adaptation of FMDV serotype O circulating in Asia. We found that the O/Cathay FMDV topotype evolves at a higher rate compared to other predominant topotypes in Asia. Genetic diversity of the O/Cathay topotype was estimated to increase in recent years, reflecting its elevated prevalence in this region. Differential host tropisms revealed the evolutionary divergence between O/Cathay and other topotypes. These findings suggest that O/Cathay FMDVs pose serious implications for the control of FMD.

# 2. Methods

## 2.1. Sequence data

Genome sequences and associated metadata of serotype O FMDV from Asia countries were collated from the GenBank database (Benson et al., 2013). Nucleotide sequences were aligned using MAFFT v7.505 (Katoh and Standley, 2013) and the VP1 protein-coding regions were extracted manually. The topotypes of all sequences were determined based on the phylogenetic proximity to reference topotype *VP1* sequences, using a neighbor-joining tree topology (Saitou and Nei, 1987) as implemented in the MEGA 11 (Tamura et al., 2021).

## 2.2. Phylodynamic reconstructions

The temporal signal of sequence data was examined in TempEst v1.5.3(Rambaut et al., 2016), using a root-to-tip regression of genetic distances against sampling time computed from the maximum-likelihood (ML) phylogenetic tree. The ML tree was inferred in RaxML v8.2.12 (Stamatakis, 2014) using 1,000 bootstrap replicates under GTR substitution model with gamma-discretized among-site rate variation, which was determined as the best-fitting nucleotide substitution model by ModelFinder (Kalyaanamoorthy et al., 2017).

Time-scaled phylogenies were reconstructed in BEAST v1.10.4(Suchard et al., 2018). The evolution of FMDV was modeled by parameterizing the process of nucleotide substitution using the GTR-gamma$_4$ model, by allowing evolutionary rates to vary across branches according to a lognormal distributed relaxed molecular clock (Drummond et al., 2006), and by using the nonparametric Skygrid coalescent demographic model (Gill et al., 2013) as tree prior, setting 100 transition-points for population size changes. The joint posterior estimates were obtained running a Markov chain Monte Carlo (MCMC) for 100–200 million iterations, 10% of which were removed as burn-in. Mixing and convergence of the MCMC chains were then assessed using Tracer v1.7.2(Rambaut et al., 2018), to ensure sufficient sampling was achieved.

## 2.3. Structural variation identification

Genome sequences of FMDV were collected and aligned pairwise to obtain mutation information. Sequences with more than 80 ‘N’ or merged nucleotides were discarded. Structural variation information of each sequence was then extracted using a Perl script. To avoid interference from sequencing quality, only sites with a gap against normal bases (i.e., A, T, C, and G) were treated as insertions or deletions.

# 3. Results

## 3.1. Cathay, ME-SA, and SEA are the predominant topotypes of FMDV serotype O circulating in Asia

To investigate the molecular epidemiology of FMDV serotype O in Asia, we compiled an extensive data set of FMDV VP1-coding sequences ($n = 3,498$) and performed phylogenetic analysis using these *VP1* gene sequences. As shown in Figure 1, at least 7 topotypes of FMDV serotype O have been found in Asia so far, comprising of Cathay, Middle East-South Asia (ME-SA), Southeast Asia (SEA), Indonesia-1, Indonesia-2, Europe-South America, and East Africa 3 (Figure 1A). Among these topotypes, Cathay, ME-SA, and SEA were the three most persistent topotypes in recent decades, with the most significant number of virus isolates. In contrast, other topotypes were only detected sporadically. Furthermore, we found that multiple topotypes were co-circulating in Asia, suggestive of complex dynamics of coexisting viral topotypes evolving within and between distinct ecological systems (Figure 1B). These findings indicate that Cathay, ME-SA, and SEA are the predominant topotypes of FMDV serotype O circulating in Asia in recent years.

## 3.2. Increased genetic diversity of Cathay topotype FMDV reflects its elevated prevalence in recent years

*VP1* sequences of Cathay, ME-SA and SEA topotype FMDV were further analyzed to reconstruct the evolutionary dynamics of these prevalent topotypes in Asia. Analyses of the root-to-tip divergence estimated from the maximum-likelihood tree as a function of the sampling time revealed strong temporal signals of both Cathay, ME-SA, and SEA FMDV evolution ($R^2$ of 0.94, 0.83, and 0.81, respectively; Figure 2A). The evolution rates were estimated to be $1.06 \times 10^{-2}$, $0.89 \times 10^{-2}$, and $0.68 \times 10^{-2}$ nucleotide substitutions/site/year for Cathay, ME-SA, and SEA topotype, respectively. These results indicate that Cathay topotype FMDV in Asia evolves at a higher rate, compared with ME-SA and SEA topotypes.

The dynamics of genetic diversity of the FMDV populations can reflect fluctuations in the size of the host population through time. We further quantified the genetic diversity of FMDV among Cathay, ME-SA, and SEA topotypes. Historical changes in viral diversity before 2006 revealed a trend of cyclical dynamics of alternating topotypes (Figure 2B). It is worth noting that, from 2011 onwards, the genetic diversity of the Cathay topotype has increased substantially, while large reductions were found in genetic diversity of both ME-SA and SEA topotypes in recent years. These results would suggest a trend that infections sustained by Cathay topotype FMDV were becoming a more severe epidemic in recent years in Asia.

## 3.3. Differential host tropisms revealed the evolutionary divergence between topotypes of serotype O FMDV in Asia

Analyzing the distributions of host species through time in the dataset, we found that the O/Cathay topotype was characterized by a highly swine-adapted tropism, as 97.9% (322/329) of O/Cathay strains were isolated from swine, with the remaining 2.1% (7/329) isolated from cattle (Figure 3A). In contrast, although the O/ME-SA topotype was able to infect many cloven-hoofed animals, including swine, cattle, and sheep, the majority of O/ME-SA isolates were obtained from cattle (88.2%), indicating a distinct host preference for O/ME-SA, compared with O/Cathay topotype (Figure 3B). The O/SEA topotype strains identified in Asia were isolated mainly from cattle until 2010, after which more isolates of the O/SEA topotype were obtained from swine, not cattle. This suggests there may be a fine-tuned tropism of the SEA topotype viruses for host species (Figure 3C).

Structural variations in the viral genome, including insertions and deletions, have a greater impact relative to nucleotide substitutions on both gene structures and protein functions, facilitating better adaptation of FMDV to hosts and/or environmental conditions. To

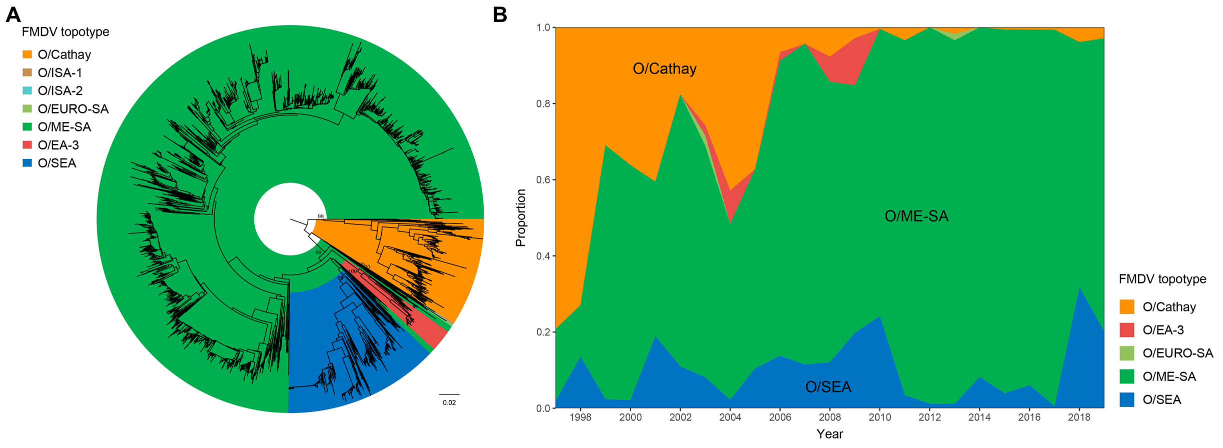


FIGURE 1
Molecular epidemiology of FMDV serotype O circulating in Asia. **(A)** Midpoint-rooted neighbor-joining tree based on VP1 gene sequences of FMDV serotype O in Asia. Node labels represent bootstrap values. Scale bar is in units of nucleotide substitutions per site. **(B)** Proportion of topotypes of FMDV serotype O in Asia between 1997 and 2019.

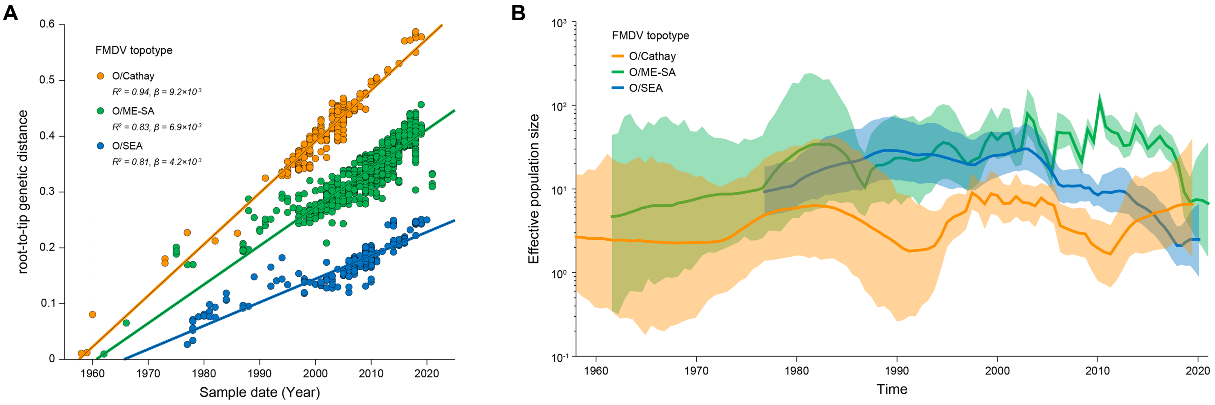


FIGURE 2
Evolutionary dynamics of dominant topotypes of serotype O FMDV circulating in Asia. **(A)** Linear regression of sampling time against divergence from the root of reconstructed maximum-likelihood trees from FMDV VP1 sequences. Circles representing tip nodes are colored according to the corresponding FMDV topotype. R-squared ($R^2$) and slope ($\beta$) parameters estimated for each fitted regression line are reported. **(B)** Historical trend of genetic diversity in dominant topotypes of serotype O FMDV circulating in Asia. Lines represent median estimates of the effective population size with colored areas defining the 95% highest posterior density region.

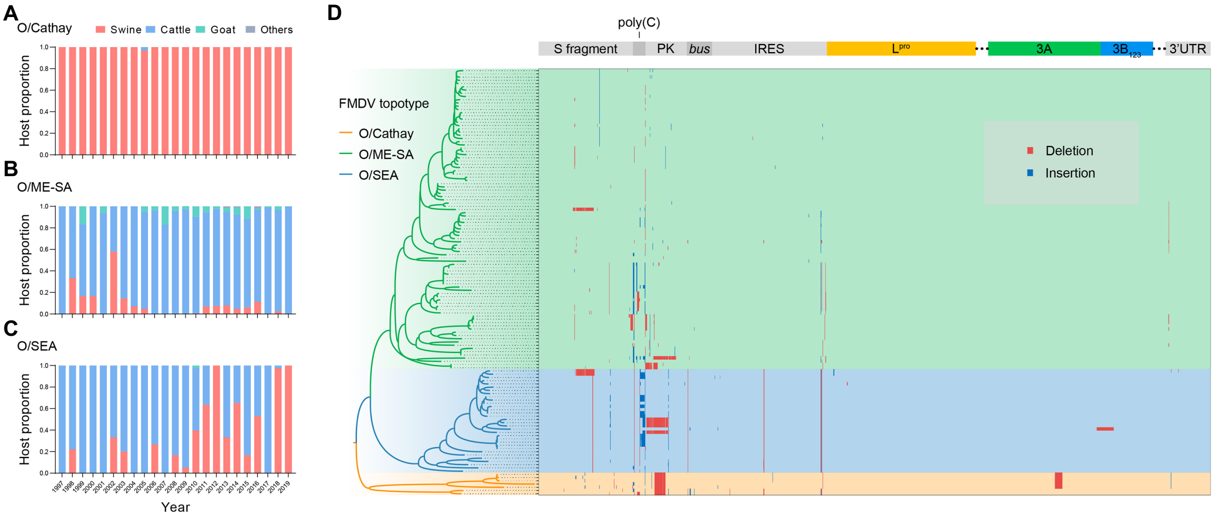


FIGURE 3
Adaptive evolution of dominant topotypes of FMDV serotype O circulating in Asia. **(A–C)**, FMDV serotype O host makeup for Cathay **(A)** ME-SA **(B)** and SEA **(C)** topotype through time. **(D)** Genomic distribution of insertions (blue) and deletions (red) in serotype O FMDV strains isolated in Asia.

further explore the potential molecular mechanism of host tropism divergence, we systematically analyzed structural variation events in whole genomes of Cathay, ME-SA, and SEA topotype FMDVs. Across the entire genome of these three topotypes, insertions and deletions occurred in the 5′ untranslated region (UTR), nonstructural proteins, and 3′ UTR (Figure 3D). It is interesting to note the deletion events observed in the pseudoknot (PK) region. A deletion of 43 nucleotides in the PK region occurred consistently in Cathay strains, and a ME-SA strain isolated from swine had a PK region deletion of 86 nucleotides. These deletions have been shown to significantly attenuate the ability to infect cattle, with no major effects on the pathogenicity in swine (Zhu et al., 2019). Furthermore, some SEA topotype strains isolated in swine in Vietnam and South Korea also included an 86-nucleotide deletion within the PK region. However, none of the genome sequences of bovine-derived SEA strains harbored this PK region deletion. These findings suggest that deletions in the PK region may reflect a common pattern of altering the host range of serotype O FMDVs.

In the 3A protein of FMDV, a 10-amino-acid deletion has been proven to be responsible for the porcinophilic properties of FMDV in previous studies (Knowles et al., 2001). This deletion was detected in recent isolates of the Cathay topotype, but not in early isolates of this topotype or isolates of the other two topotypes (Figure 3D). In addition, a dual structural variation, a 70-nucleotide deletion in the S fragment combined with a 1-amino-acid insertion in the leader protein ($L^{pro}$), which was demonstrated as a determinant of attenuated virulence of serotype O FMDVs in cattle (Yang et al., 2020), was observed in two SEA topotype strains isolated from swine (Figure 3D). These facts indicate that the divergence of host tropism may be due to accumulated structural variations across the viral genome, rather than a single indel mutation.

## 4. Discussion

In this study, we describe the phylogeny and evolution of serotype O FMDV in Asia. Our findings indicate that Cathay, ME-SA, and SEA are the predominant topotypes of FMDV serotype O circulating in Asia in recent years, and suggest a trend that infections sustained by Cathay topotype FMDV have become dominant in recent years in Asia. Analyses of evolutionary divergence between topotypes of serotype O FMDV highlighted the significant role of accumulated structural variations across the viral genome in the divergence of host tropism.

FMD is a global disease, which poses a major threat to the animal industry and causes enormous economic losses (Porphyre et al., 2018). Among FMDV serotypes, serotype O is most prevalent and the maximum outbreaks occurred with this serotype in Asian countries (Hossen et al., 2020). In recent years, Cathay, ME-SA, and SEA are the predominant topotypes of FMDV serotype O circulating in Asia. So far, cattle infected with FMDV of the O/Cathay topotype have a low risk of viral transmission or persistence, which is a major reason for the smaller number of outbreaks caused by this topotype compared to others (Nishi et al., 2021). However, sporadic outbreaks continue to be reported in several Southeast Asian countries. In previous studies, where and how viruses of this topotype are maintained or spread remain unclear (Di Nardo et al., 2014; Brito et al., 2017). In our research, we show that the genetic diversity of the Cathay topotype has increased substantially characterized by a highly swine-adapted tropism, causing continuous prevalence of O/Cathay. Because of low antigenic matching between the O/Cathay strain and current vaccine strains according to quarterly reports from the World Reference Laboratory for FMD[1], strategies to eradicate FMDV of O/Cathay would rely on movement restriction and culling of animals suspected of infection (Nishi et al., 2021). The pig industry is one of the most important sectors of agriculture in most countries of Asia, where such strategies can lead to severe economic losses. Further statistical surveillance should be targeted toward O/Cathay, a kind of porcinophillic FMDV, to strategize appropriate risk management and to reduce the possibility of virus transmission.

The genetic variations in Cathay topotype viruses have accumulated over several decades. In previous research, the deletion of 43 nt in the PK region, the 10-amino-acid deletion in the 3A protein and the 70-nt deletion in the S fragment or the single leucine insertion in $L^{pro}$ of serotype O FMDV may show a swine-adapted characteristic, resulting in the altered host tropism of the virus in cattle (Knowles et al., 2001; Zhu et al., 2019; Yang et al., 2020). Our findings in this study further indicated the important role of PK region deletion in the variation and the accumulated structural variations across the viral genome for the host tropism of serotype O FMDVs, which might be the critical determinants of viral tropism of serotype O FMDV from cattle to swine. The concurrence of these mutations in serotype O FMDV may result in the altered host range of the virus that enabled swine to become the main epidemiological host.

The serotype O FMDV is the most prevalent serotype in Asia. Asia possesses a dense pig population, which probably caused the more frequent propagation and prevalence of the swine-origin serotype O FMDV. Besides, the decreased viral pathogenicity of swine-adapted FMDV may benefit virus maintenance in the pigs, because high pathogenicity may leave the host unable to further support viral maintenance or reproduction (Ebert and Bull, 2003; Dortmans et al., 2010; Wang et al., 2012). Thus, continual systematic surveillance and more detailed investigation of prevailing serotype O FMDVs in swine populations, especially the O/Cathay topotype, are urgently needed to formulate an efficient FMD control strategy for Asia.

## Data availability statement

The datasets presented in this study can be found in the GenBank database (https://www.ncbi.nlm.nih.gov/genbank/). The accession numbers can be found in the Supplementary material.

## Author contributions

FL, QZ, and YZ contributed to the conception and design of the study. YLi and JM organized the database. FL and RW performed the statistical analysis. FL wrote the first draft of the manuscript. YLiu and XZ wrote sections of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

## Funding

This work was supported by grants from the National Key R&D Program of China (2021YFD1800300) and National Natural Science Foundation of China (32202791).

## Acknowledgments

We thank Lele Zhao for comments on an earlier version 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.

## 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.2023.1147652/full#supplementary-material

[1] https://www.wrlfmd.org/ref-lab-reports

## References

Adhikari, G., Acharya, K. P., Upadhyay, M., Raut, R., Kaphle, K., Khanal, T., et al. (2018). Outbreak investigations of foot and mouth disease virus in Nepal between 2010 and 2015 in the context of historical serotype occurrence. *Vet. Med. Sci.* 4, 304–314. doi: 10.1002/vms3.120

Beard, C. W., and Mason, P. W. (2000). Genetic determinants of altered virulence of Taiwanese foot-and-mouth disease virus. *J. Virol.* 74, 987–991. doi: 10.1128/jvi.74.2.987-991.2000

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi: 10.1093/nar/gks1195

Blacksell, S. D., Siengsanan-Lamont, J., Kamolsiripichaiporn, S., Gleeson, L. J., and Windsor, P. A. (2019). A history of FMD research and control programmes in Southeast Asia: lessons from the past informing the future. *Epidemiol. Infect.* 147:e171. doi: 10.1017/s0950268819000578

Brito, B. P., Rodriguez, L. L., Hammond, J. M., Pinto, J., and Perez, A. M. (2017). Review of the global distribution of foot-and-mouth disease virus from 2007 to 2014. *Transbound. Emerg. Dis.* 64, 316–332. doi: 10.1111/tbed.12373

Di Nardo, A., Knowles, N. J., Wadsworth, J., Haydon, D. T., and King, D. P. (2014). Phylodynamic reconstruction of O CATHAY topotype foot-and-mouth disease virus epidemics in the Philippines. *Vet. Res.* 45:90. doi: 10.1186/s13567-014-0090-y

Diaz-San Segundo, F., Medina, G. N., Stenfeldt, C., Arzt, J., and de Los Santos, T. (2017). Foot-and-mouth disease vaccines. *Vet. Microbiol.* 206, 102–112. doi: 10.1016/j.vetmic.2016.12.018

Doel, T. R. (2003). FMD vaccines. *Virus Res.* 91, 81–99. doi: 10.1016/s0168-1702(02)00261-7

Dortmans, J. C., Rottier, P. J., Koch, G., and Peeters, B. P. (2010). The viral replication complex is associated with the virulence of Newcastle disease virus. *J. Virol.* 84, 10113–10120. doi: 10.1128/jvi.00097-10

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88. doi: 10.1371/journal.pbio.0040088

Dunn, C. S., and Donaldson, A. I. (1997). Natural adaption to pigs of a Taiwanese isolate of foot-and-mouth disease virus. *Vet. Rec.* 141, 174–175. doi: 10.1136/vr.141.7.174

Ebert, D., and Bull, J. J. (2003). Challenging the trade-off model for the evolution of virulence: is virulence management feasible? *Trends Microbiol.* 11, 15–20. doi: 10.1016/s0966-842x(02)00003-3

Freimanis, G. L., Di Nardo, A., Bankowska, K., King, D. J., Wadsworth, J., Knowles, N. J., et al. (2016). Genomics and outbreaks: foot and mouth disease. *Rev. Sci. Tech.* 35, 175–189. doi: 10.20506/rst.35.1.2426

Garcia, M. (2012). *Viral Genomes-Molecular Structure, Diversity, Gene Expression Mechanisms and Host-Virus interactions*. Rijeka: IntechOpen.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Hossen, M. L., Ahmed, S., Khan, M. F. R., Nazmul Hussain Nazir, K. H. M., Saha, S., Islam, M. A., et al. (2020). The emergence of foot-and-mouth disease virus serotype O PanAsia-02 sub-lineage of Middle East-south Asian topotype in Bangladesh. *J. Adv. Vet. Anim. Res.* 7, 360–366. doi: 10.5455/javar.2020.g429

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

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

Knowles, N. J., Davies, P. R., Henry, T., O'Donnell, V., Pacheco, J. M., and Mason, P. W. (2001). Emergence in Asia of foot-and-mouth disease viruses with altered host range: characterization of alterations in the 3A protein. *J. Virol.* 75, 1551–1556. doi: 10.1128/jvi.75.3.1551-1556.2001

Knowles, N. J., and Samuel, A. R. (2003). Molecular epidemiology of foot-and-mouth disease virus. *Virus Res.* 91, 65–80. doi: 10.1016/s0168-1702(02)00260-5

Lee, G., Hwang, J. H., Park, J. H., Lee, M. J., Kim, B., and Kim, S. M. (2020). Vaccine strain of O/ME-SA/Ind-2001e of foot-and-mouth disease virus provides high immunogenicity and broad antigenic coverage. *Antivir. Res.* 182:104920. doi: 10.1016/j.antiviral.2020.104920

Li, P., Huang, S., Zha, J., Sun, P., Li, D., Bao, H., et al. (2022). Evaluation of immunogenicity and cross-reactive responses of vaccines prepared from two chimeric serotype O foot-and-mouth disease viruses in pigs and cattle. *Vet. Res.* 53:56. doi: 10.1186/s13567-022-01072-7

Li, K., Wang, C., Yang, F., Cao, W., Zhu, Z., and Zheng, H. (2021). Virus-host interactions in foot-and-mouth disease virus infection. *Front. Immunol.* 12:571509. doi: 10.3389/fimmu.2021.571509

Mahapatra, M., Upadhyaya, S., Aviso, S., Babu, A., Hutchings, G., and Parida, S. (2017). Selection of vaccine strains for serotype O foot-and-mouth disease viruses (2007-2012) circulating in Southeast Asia, East Asia and Far East. *Vaccine* 35, 7147–7153. doi: 10.1016/j.vaccine.2017.10.099

Nishi, T., Morioka, K., Kawaguchi, R., Yamada, M., Ikezawa, M., and Fukai, K. (2021). Quantitative analysis of infection dynamics of foot-and-mouth disease virus strain O/CATHAY in pigs and cattle. *PLoS One* 16:e0245781. doi: 10.1371/journal.pone.0245781

Park, S. H., Lee, S. Y., Kim, J. S., Kim, A. Y., Park, S. Y., Lee, J. H., et al. (2021). Scale-up production of type O and a foot-and-mouth disease bivalent vaccine and its protective efficacy in pigs. *Vaccines (Basel)* 9:586. doi: 10.3390/vaccines9060586

Porphyre, T., Rich, K. M., and Auty, H. K. (2018). Assessing the economic impact of vaccine availability when controlling foot and mouth disease outbreaks. *Front. Vet. Sci.* 5:47. doi: 10.3389/fvets.2018.00047

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian Phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly path-O-gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

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

Shao, J. J., Wong, C. K., Lin, T., Lee, S. K., Cong, G. Z., Sin, F. W., et al. (2011). Promising multiple-epitope recombinant vaccine against foot-and-mouth disease virus type O in swine. *Clin. Vaccine Immunol.* 18, 143–149. doi: 10.1128/cvi.00236-10

Sinkala, Y., Simuunza, M., Pfeiffer, D. U., Munang'andu, H. M., Mulumba, M., Kasanga, C. J., et al. (2014). Challenges and economic implications in the control of foot and mouth disease in sub-saharan Africa: lessons from the Zambian experience. *Vet. Med. Int.* 2014:373921. doi: 10.1155/2014/373921

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

Subramaniam, S., Mohapatra, J. K., Sahoo, N. R., Sahoo, A. P., Dahiya, S. S., Rout, M., et al. (2022). Foot-and-mouth disease status in India during the second decade of the twenty-first century (2011-2020). *Vet. Res. Commun.* 46, 1011–1022. doi: 10.1007/s11259-022-10010-z

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

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

Upadhyaya, S., Mahapatra, M., Mioulet, V., and Parida, S. (2021). Molecular basis of antigenic drift in serotype O foot-and-mouth disease viruses (2013-2018) from Southeast Asia. *Viruses* 13:1886. doi: 10.3390/v13091886

Wang, Q., Gao, Y., Wang, Y., Qin, L., Qi, X., Qu, Y., et al. (2012). A 205-nucleotide deletion in the 3′ untranslated region of avian leukosis virus subgroup J, currently emergent in China, contributes to its pathogenicity. *J. Virol.* 86, 12849–12860. doi: 10.1128/jvi.01113-12

Yang, F., Zhu, Z., Cao, W., Liu, H., Wei, T., Zheng, M., et al. (2020). Genetic determinants of altered virulence of type O foot-and-mouth disease virus. *J. Virol.* 94:e01657-19. doi: 10.1128/jvi.01657-19

Zhu, Z., Yang, F., Cao, W., Liu, H., Zhang, K., Tian, H., et al. (2019). The pseudoknot region of the 5' untranslated region is a determinant of viral tropism and virulence of foot-and-mouth disease virus. *J. Virol.* 93:e02039-18. doi: 10.1128/jvi.02039-18

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fmicb.2023.1136386

# Variation in synonymous evolutionary rates in the SARS-CoV-2 genome

Qianru Sun[1,2], Jinfeng Zeng[1,2], Kang Tang[1,2], Haoyu Long[1,2], Chi Zhang[1,2], Jie Zhang[1,2], Jing Tang[1,2], Yuting Xin[1,2], Jialu Zheng[1,2], Litao Sun[1,2], Siyang Liu[1,2] and Xiangjun Du[1,2,3]*

[1]School of Public Health (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Shenzhen, China, [2]School of Public Health (Shenzhen), Sun Yat-sen University, Guangzhou, China, [3]Key Laboratory of Tropical Disease Control, Ministry of Education, Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

EDITED BY
Zhan Zhou,
Zhejiang University, China

REVIEWED BY
Zhenglin Zhu,
Chongqing University, China
Denis Jacob Machado,
University of North Carolina at Charlotte, United States

*CORRESPONDENCE
Xiangjun Du
✉ duxj9@mail.sysu.edu.cn

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 03 January 2023
ACCEPTED 13 February 2023
PUBLISHED 09 March 2023

CITATION
Sun Q, Zeng J, Tang K, Long H, Zhang C, Zhang J, Tang J, Xin Y, Zheng J, Sun L, Liu S and Du X (2023) Variation in synonymous evolutionary rates in the SARS-CoV-2 genome. *Front. Microbiol.* 14:1136386. doi: 10.3389/fmicb.2023.1136386

**Introduction:** Coronavirus disease 2019 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Influential variants and mutants of this virus continue to emerge, and more effective virus-related information is urgently required for identifying and predicting new mutants. According to earlier reports, synonymous substitutions were considered phenotypically silent; thus, such mutations were frequently ignored in studies of viral mutations because they did not directly cause amino acid changes. However, recent studies have shown that synonymous substitutions are not completely silent, and their patterns and potential functional correlations should thus be delineated for better control of the pandemic.

**Methods:** In this study, we estimated the synonymous evolutionary rate (SER) across the SARS-CoV-2 genome and used it to infer the relationship between the viral RNA and host protein. We also assessed the patterns of characteristic mutations found in different viral lineages.

**Results:** We found that the SER varies across the genome and that the variation is primarily influenced by codon-related factors. Moreover, the conserved motifs identified based on the SER were found to be related to host RNA transport and regulation. Importantly, the majority of the existing fixed-characteristic mutations for five important virus lineages (Alpha, Beta, Gamma, Delta, and Omicron) were significantly enriched in partially constrained regions.

**Discussion:** Taken together, our results provide unique information on the evolutionary and functional dynamics of SARS-CoV-2 based on synonymous mutations and offer potentially useful information for better control of the SARS-CoV-2 pandemic.

KEYWORDS

**binding motif, codon usage, dominant variants, SARS-CoV-2, synonymous evolutionary rate**

## Introduction

Since its first appearance 3 years ago, coronavirus disease 2019, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been declared a global pandemic, and influential variants continue to emerge and spread globally. For better monitoring and research (Tao et al., 2021; Kumar et al., 2022), the World Health Organization has listed some of the key viral variants or lineages with important mutations as variants of concern. Variants with different mutation combinations can emerge within short periods and have different effects. Therefore, it is crucial to understand this process from the evolutionary perspective for better prevention and control of the epidemic. Currently, whether a mutation is deleterious is primarily determined by comparing the relevant lineage with reference sequences through multiple sequence alignment or evidence from biological experiments (Badua et al., 2021; Lauring and Hodcroft, 2021). For example, Nextstrain uses the number of mutations at each site or the entropy of change to represent the site's degree of variability based on phylogenetic trees and some viral infection experiments assessing specific mutations (Hadfield et al., 2018; Zhang L. et al., 2020; Motozono et al., 2021; Tao et al., 2021). However, whether new mutants truly increase virus transmissibility and infectivity depends not only on the accumulation of mutations but also on the recurrence or rapid removal of mutations and their epistatic effects. Traditional analytical methods based on fixed mutations can elucidate the importance of mutations; however, owing to the time-consuming experimental verification and the rapidity of viral mutations, new methods are warranted for better and timely acquisition of updated critical information.

The *dN/dS* ($\mathrm{K_a/K_s}$) value, where *dN* or $K_a$ represents the number of non-synonymous substitutions/number of non-synonymous sites and *dS* or $K_s$ represents the number of synonymous substitutions/number of synonymous sites, is always used to determine whether there is evidence for the selection of species, lineages, or proteins and gene areas (Duffy et al., 2008; Wilson et al., 2020; MacLean et al., 2021). In reality, the majority of the observed mutations are a result of natural selection and genetic drift. The aforementioned *dN/dS* indicator can also be used to determine the direction of selection. *dN* is more impacted by natural selection because amino acid alterations are always generated through selection; by contrast, *dS* is more related to the background mutation rates because such mutations do not directly cause amino acid changes. However, whether synonymous mutations represent the complete viral background remains a matter of debate in recent years. Some studies have suggested that a substantial proportion of synonymous alterations are not silent; selection, codon usage, and other factors can influence synonymous variations (de Oliveira et al., 2021; Mordstein et al., 2021; Rahman et al., 2021; Shen et al., 2022). However, it remains unclear how data on synonymous mutations in the SARS-CoV-2 genome can offer additional, in-depth knowledge on evolutionary processes and inform rules and guidelines for the precise prevention and control of the pandemic.

Furthermore, a viral infection of host cells is a complex, multistep, and often specific process. Like other RNA viruses, SARS-CoV-2 relies on regulators to effectively utilize host cellular factors at many biochemical levels, including RNA stability, processing, localization, and translation, to facilitate replication and progeny production (Flynn et al., 2021). Although existing studies have explored the proteins that can bind viral RNA and their downstream regulatory metabolic pathways from the host's perspective (Flynn et al., 2021; Khan et al., 2021; Schmidt et al., 2021), the viral genome is known to mutate faster than the host genome. This feature jeopardizes the efficacy of vaccines and drugs. Moreover, different regions of viral genomes evolve at different rates, with some regions being hypervariable and others being conserved. Until now, few studies have assessed the conservation of the virus and its relationship with the interaction patterns between viruses and hosts, especially from the perspective of synonymous mutations; more studies are needed to explore this further.

Based on the foregoing questions, it is important to explore the synonymous evolutionary rate (SER) in the open reading frames (ORFs) of the SARS-CoV-2 genome, the factors that influence the SER, and what rules can be drawn through comparison of fixed-characteristic amino acid mutations with different lineages. To answer these questions, using a mutation network approach (Zhang C. et al., 2020; Wang Y. et al., 2021), we described the distribution of the SER across the SARS-CoV-2 genome along with its influential factors and explored the conserved motifs based on the SER and the motifs' potential functional relationships with the host by performing enrichment analyses. We also assessed the potentially important and functional amino acid mutations based on the SER for identifying future dominant variants to better control the pandemic.

# Materials and methods

## Sequence data

A total of 2,537,286 original SARS-CoV-2 genomic sequences were downloaded from the Global Initiative on Sharing All Influenza Data system (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare et al., 2021) as of 15 September 2021. Sequences were excluded if they met any of the following criteria: (1) genome size of <29,000 nucleotides; (2) >5% of undetermined nucleotides; (3) non-human host. To further ensure sequence quality, sequences with complete collection date, region details (specific to the country), and a gap length of <400 bp were included. The sequences were first aligned using MAFFT v7.310 (Katoh and Standley, 2013), with Wuhan-Hu-1 (MN908947.3) as the reference. The alignment command was as follows: *mafft-—6merpair—thread-12—keeplength—addfragments othersequences referencesequence > output*. Moreover, the redundant sequences, which are sequences with identical nucleotide compositions, were filtered out; however, the redundant sequence with the earliest collection time was included because the connected edges of the mutation network are based on the mutation probability. If two sequences were the same (without any mutation), the probability between them was 1. Therefore, we believed that only transmission and no evolution occurred between the two sequences and that they could not provide more evolutionary information. Next, we conducted stratified sampling per country (region) per day. Finally, a total of 10,089 sequences were included in this study (accessible at 10.55876/gis8.230130ru; also, in Supplementary Table 7). We also masked the problematic sites to avoid artificial errors using the methods outlined at https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480 (Oliver et al., 2021). Finally, we filled the undetermined nucleotides or gaps with the element with the highest frequency at the corresponding position based on the top 10 closest sequences measured by the Hamming distance (Wang Y. et al., 2021). Subsequently, except for stop codons and non-coding sites, sites corresponding to protein-coding ORFs were mapped to the reference sequence alignment and eventually used to construct the mutation network.

## SER estimation

Following the methods outlined by Zhang C. et al. (2020), a directed and weighted mutation network was constructed with the nodes representing strains and links which represent pairs of strains, and a mutation probability of no more than the predetermined threshold (>10th percentile). The baseline mutation probabilities among A, T, G, and C were extracted from pairs of sequences with single-nucleotide differences in the corresponding data. The mutation probabilities between pairs of strains with different numbers of mutations were calculated as the product of probabilities of the single mutations (Zhang C. et al., 2020; Wang Y. et al., 2021) (Supplementary Figure 1). Paths on the network were extracted using the random walk method. First, 20,000 start nodes were randomly chosen that have descendants (out-degree $\neq$ 0), and second, a random walk was executed from selected start nodes. The paths yielded by the repeated random walk were considered evolutionary paths in the real world. Moreover, the nodes on the path had an evolutionary ancestor–descendant relationship. To ensure sufficient divergence, only paths with more than 1 month were included. Python package NetworkX v2.8.4 was used for the analysis (Hagberg et al., 2008). In summary, for the final mutation network, the input was 10,089 sequences and the output was the evolutionary paths got from random walks.

Different ORFs in the SARS-CoV-2 genome have different lengths (Supplementary Table S1). To avoid biases caused by the ORF length, we used a codon-based sliding window approach; a 600-bp window and 3-bp step were maintained. The 600-bp window was set after considering the upper limit of the substitution rate of the virus to ensure sufficient observation of substitutions along any chosen path. The KaKs_Calculator v2.0 software MLWL model was used to calculate the *dS* value (Tzeng et al., 2004; Wang et al., 2010). The following command was used: *KaKs_Calculator -i input -o out -m MLWL*. Next, a linear regression analysis of *dS* on the collection time interval was performed, and the regression line slope was represented as the SER for the start position of the window (Ho and Duchene, 2014; Kim et al., 2022). The Kruskal–Wallis test and Mann–Whitney U-test were used to compare the statistical differences between the ORFs. Based on the SER (10th percentile, 50th percentile, and 90th percentile), the genome was divided into four regions: (1) the free region (the region with an upper 90th percentile SER); (2) the slightly free region (the region with an SER between the 50th percentile and 90th percentile); (3) the partially constrained region (the region with an SER between the 10th percentile and 50th percentile); and (4) the constrained region (the region with an SER lower than the 10th percentile SER).

## Motif identification and function association analysis

With the constrained regions set as the target and three other groups set as the background, we used STREME v5.5.0 and a zero-order Markov model for background model creation in the MEME suite server to find conserved sequence patterns (motifs) with a sequence length of 3–30 bp (Bailey et al., 2015; Bailey, 2021), a *P*-value of <0.001, and coverage of >70%. Next, we used the find individual motif occurrence (FIMO v5.5.0) program to locate the motif position with a *P*-value of <1e−4 for double chains in the sequence (Grant et al., 2011).

The RNA motif data recognized by RNA-binding proteins (RBPs) were obtained from a previous study (Ray et al., 2013); only records from *Homo sapiens* were included. The Tomtom motif comparison tool v5.5.0 in the MEME suite server is used to compare motifs against a database of known motifs. In this study, we used this tool to compare motif similarity and identify host-associated proteins with default settings (Gupta et al., 2007). Cytoscape v3.8.0 was used to visualize the protein–motif relationships (Shannon et al., 2003). We also conducted Gene Ontology (GO) enrichment analysis based on the hypergeometric distribution using clusterProfiler v4.6.0 package in R with default parameters (Yu et al., 2012).

## Feature collection and model construction

To determine the dinucleotide composition (CpG and UpA), we divided the dinucleotide frequency within the sequence by the product of the frequency of each nucleotide (Mordstein et al., 2021). All codon usage index types, including the codon bias index, the effective number of codons, GC content and GC content in the third codon (GC, GC3), and silent base composition (A3, T3, G3, and C3), were calculated using CodonW v1.4.4 with default parameters (Peden, 2000); the protein hydrophobicity was also calculated using CodonW. The ω (*dN/dS*) value, which represents the selection of entire ORFs, was estimated using the BUSTED method in HyPhy v2.5.2 with default parameters (Murrell et al., 2015). By contrast, the non-synonymous evolutionary rate (NER), which represents the selection in codon sites, was calculated similarly to SER by fitting the regression line of *dN* and the collection time interval. The normalized van der Waals volume and relative mutability for each window were extracted and calculated using the AAindex2 database (Kawashima et al., 2008). The minimum free energy of the RNA secondary structure in the windows was determined using RNAstructure Fold server v6.4 with the default parameters (Reuter and Mathews, 2010). Based on the absolute difference between the two sequences, the aforementioned features were used for the following analysis: for motif information, "0" was assigned if the motif did not exist; "1" was assigned if the motif existed in one sequence; and "2" was assigned if the motif existed in both sequences.

Features were filtered based on the results of Spearman's correlation analysis. Based on the aforementioned features, a light gradient-boosting machine (LightGBM) regression model was constructed for determining the SER, and R-squared values were used to measure any explicable variations (Meng and Liu, 2017). Next, 80% of the randomly selected data were used as the training set, and the remaining 20% were set as the test set. The GridSearchCV technique and 10-fold cross-validation were employed to determine the best hyperparameters for model construction (Pedregosa et al., 2011). Subsequently, the SHapley Additive exPlanations (SHAP) value was used to explain the output of the constructed machine learning model to evaluate feature importance (Lundberg et al., 2020). The feature value represents the value of each feature in the model, ranging from small to large and from blue color to red. The SHAP value represents the direction and size of the SER affected by each sample; a value >0 indicates a positive impact, and any other value indicates a negative impact. LightGBM v3.3.3, scikit-learn v1.0.2, and shap v0.41.0 packages were used for these analyses.

## Comparison of fixed-characteristic mutations in different lineages

Fixed-characteristic amino acid mutations, including deletions accumulated in different lineages, were downloaded from the Cov-Lineages repository (https://cov-lineages.org/lineage_list.html). Characteristic mutations in the lineages Alpha, Beta, Gamma, Delta, and Omicron (sub-lineages: BA.1, BA.2, BA.2.12.1, BA.2.75, BA.4, and BA.5) were used in our analysis (Supplementary Table 6; Figure 4C).

## Statistical analysis

The Kruskal–Wallis, Mann–Whitney U, and chi-square tests ($\alpha = 0.05$) were used with the stats.kruskal function, stats.mannwhitneyu function, and stats.chi2_contigency function, respectively, in SciPy 1.5.2 package in Python 3.8.5. Furthermore, the ggplot2 3.3.5 package in R 4.1.1 and matplotlib 3.3.2 in Python 3.8.5 were used to generate most figures.

# Results

## SER landscape for the SARS-CoV-2 genome

We constructed the mutation network such that it was scale-free (Supplementary Figure 2). Based on the created mutation network, random walks were executed 20,000 times, and the potential paths between sequence pairs were extracted. Because of the strong similarities among SARS-CoV-2 viruses, only paths between paired nodes with a time interval of >1 month were included in the following analysis.

In general, the SER distribution across the whole genome was extremely skewed and lopsided, displaying the characteristics of Gamma distribution, with a median (Q1, Q3) of $6 \times 10^{-4}$ ($4 \times 10^{-4}$, $1.1 \times 10^{-3}$) per site per year across all regions (Figure 1A). The SER was highly variable, with averages ranging from $5 \times 10^{-4}$ to $2 \times 10^{-3}$ per site per year (Figure 1B). Moreover, the SERs of different ORFs ($H = 982.1478$, $P < 0.001$) and between any of the ORFs (adjusted $P < 0.05$) were significantly different. The SERs within the SARS-CoV-2 genome were also substantially different (Figure 1C). The fluctuations were obvious, as indicated through traditional diversity cues, implying that the SERs varied widely and the synonymous substitutions tended to be enriched or reduced in specific genomic regions. Based on the SERs (10th percentile, 50th percentile, and 90th percentile), the genome was divided into four regions, as explained in the Methods section (Figures 1A, D). The overall SER for the S gene was low and mostly located within the partly constrained region (Figure 1C), which was different from that identified in the traditional diversity analysis (Supplementary Figure 3). This difference was not caused by the increased NER (Supplementary Figure 4). Moreover, the SER in the ORF1ab region tended to have more freedom toward a greater variation.

## Characteristics of the conserved motifs in the constrained region

To check whether conserved sequences (motifs) existed in the constrained region (Supplementary Table 2), we performed an enrichment analysis for comparing sequences in the constrained region using other regions as the background. After strict filtering, we obtained 10 motifs with a length of ∼9–15 bp (Figure 2A; Supplementary Table 8). The Kruskal–Wallis test results indicated that the base composition was statistically significant and that the A + T content in the motifs was higher than the G + C content ($P = 6e{-4}$) (Figure 2B). Furthermore, these motifs were found in various ORFs throughout the genome (Figure 2C).

Previous studies have revealed that some regions of the viral genome are preferred by host proteins (Flynn et al., 2021; Khan et al., 2021; Lu et al., 2021; Schmidt et al., 2021). In other words, the host RBPs could specifically bind certain sequences such as motifs on the viral genome. The identified motifs from the viral genome were thus compared with some known binding motifs of the host RBPs. A total of 30 host protein genes were found to be associated with the 10 identified motifs (Figure 2D). Of note, some motifs may be targeted by more than one host protein, and the same host protein may bind different motifs in the viral genome. Remarkably, *YBX1*, which was identified to bind Motifs3 and Motif6, was found to be associated with viral infections, including SARS-CoV-2 and Zika, and previous experiments have shown that knockout of this gene can reduce the infection intensity (Zhang et al., 2022). Some other associated host proteins were also found in some experimental studies assessing viral infection; for example, *SFPQ* was found to interact with the SARS-CoV-2 genome and promote viral RNA amplification (Labeau et al., 2022). Functional GO annotation revealed that these genes are involved in metabolic RNA regulation (Figure 2E).

## Factors contributing to the SER variations

To further investigate the factors that may contribute to the variations in the SERs in the SARS-CoV-2 genome, the codon usage index, the dinucleotide composition, the selection index, the structure index, and the motif information were included and fed into the model. The features were classified into five groups: the codon usage index, selection index, dinucleotide composition, structure index (Resch et al., 2007; Callens et al., 2021; McGrath, 2021; Mordstein et al., 2021; Pintó and Bosch, 2021), and conserved motifs were identified in this study (Supplementary Table 3). G3, gravy, *van der Waals* volume, and aa mutations were excluded owing to high collinearity based on the correlation coefficients ($R^2 > 0.9$, Supplementary Figure 5),

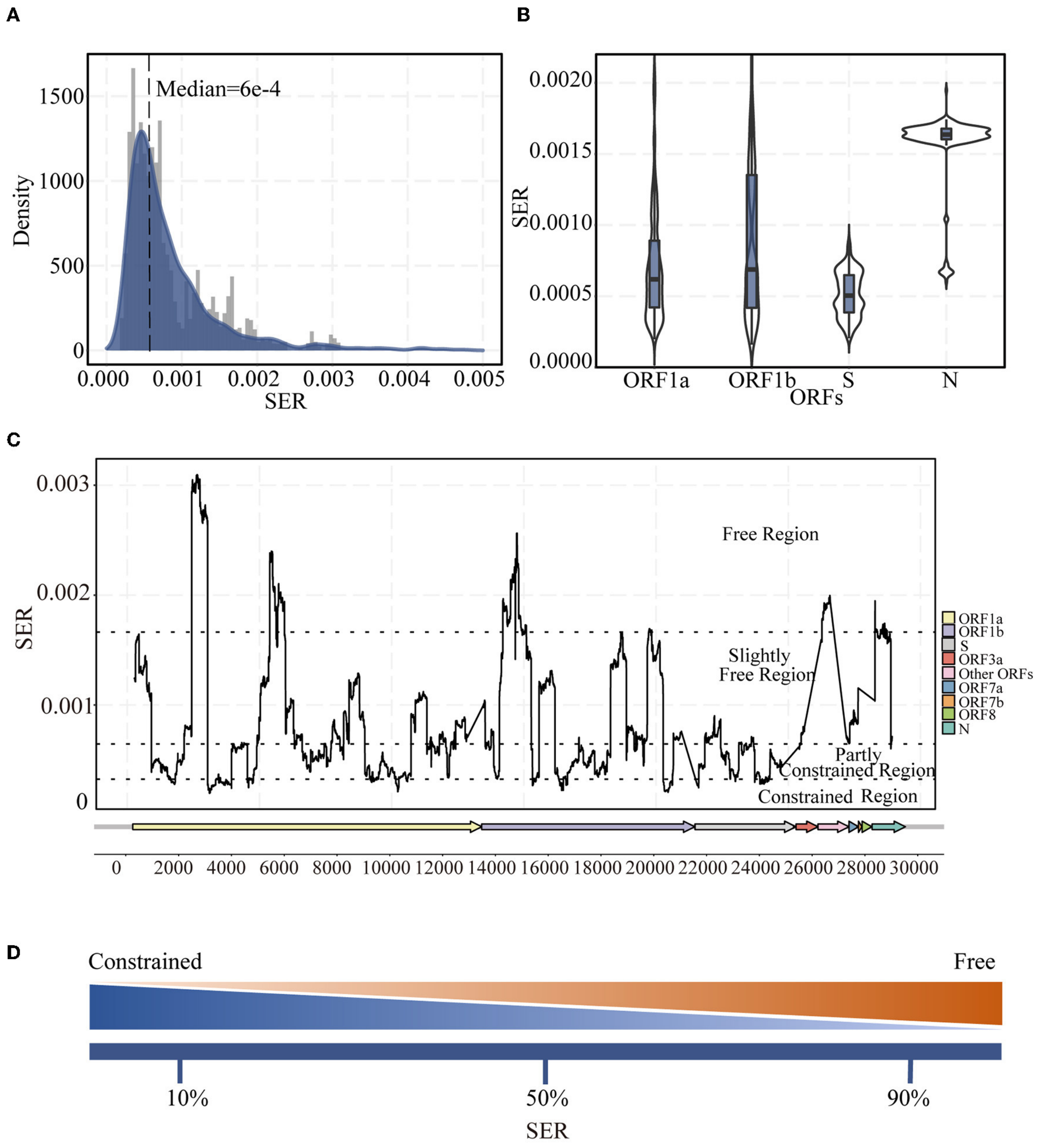


FIGURE 1
Landscape of synonymous evolutionary rate (SER) of the SARS-CoV-2 genome. **(A)** SER density distribution in all ORFs of SARS-CoV-2. **(B)** Violin plot of SER distribution for representative ORF1a, ORF1b, S, and N regions. **(C)** SER across the whole genomes based on sliding windows. Black dotted lines were 90th, 50th, and 10th percentile levels of SER. **(D)** Percentiles are used to divide regions. The greater the SER, the more freedom; the smaller the SER, the greater the constraint.

whereas the other features were included and used in the LightGBM model.

Based on cross-validation, the best model after grid search (Supplementary Table 4) had an adjusted $R^2$ of 0.72 on the training dataset and 0.69 on the test dataset, indicating good performance (Supplementary Table 5). According to the final model, factors from the codon usage index group contributed the most to the variations in the SERs (80.37%). GC3 (36.32%) was the most important single feature, followed by the non-SER (16.60%) from the group of selection (Figure 3).

## Association between the accumulated characteristic mutations and SERs

The characteristic mutations accumulated in the five main lineages (Alpha, Beta, Gamma, Delta, and Omicron) were mapped onto the SER landscape of the SARS-CoV-2 genome to investigate their associations (Figure 4). Based on the classification of the four regions across the genome based on the SER landscape, because most mutations exist in the middle region, the chi-square test was used to compare the number of characteristic mutations between

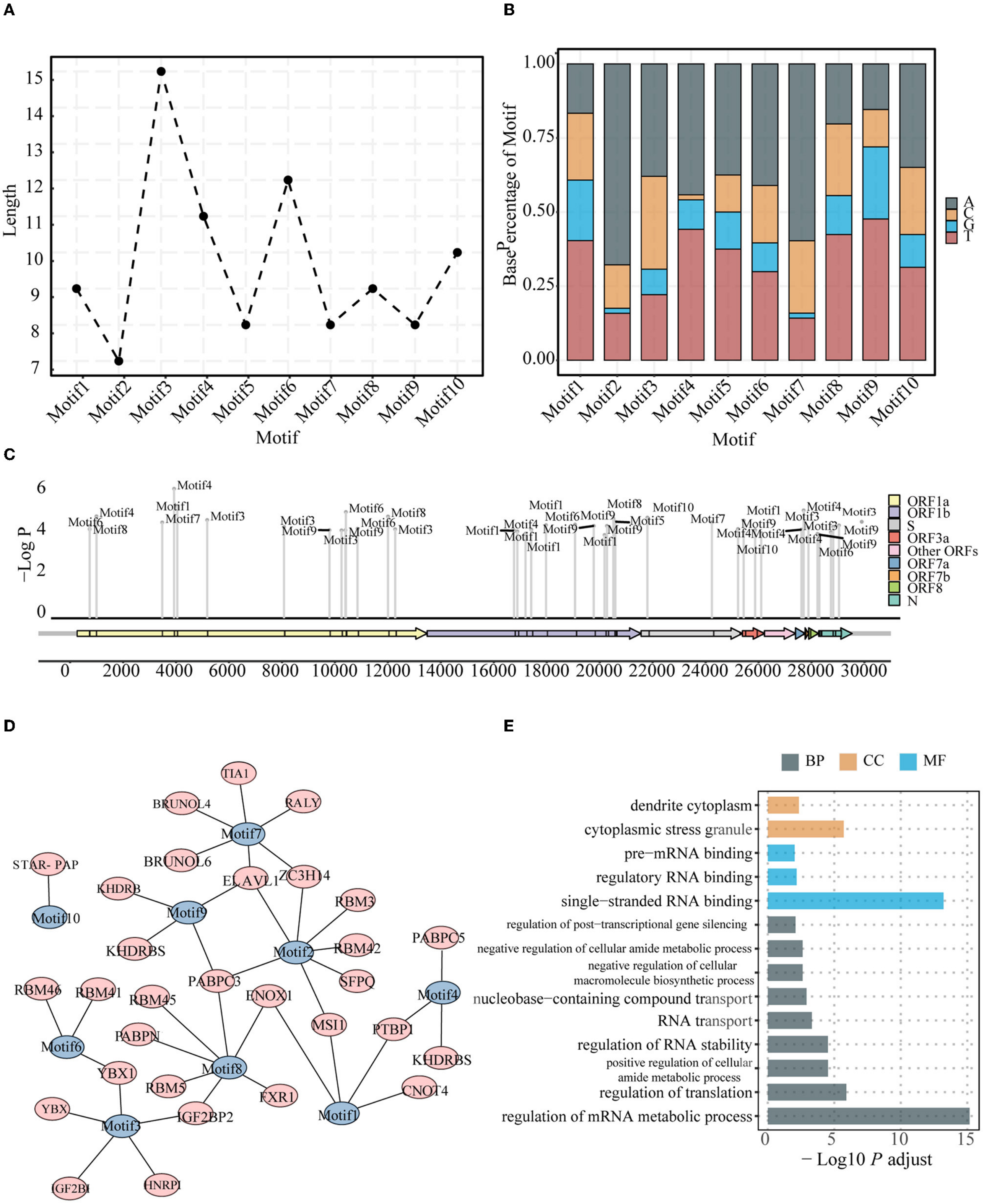


FIGURE 2

Characteristics for the conserved motifs. Length **(A)** and Base composition **(B)** of identified motifs enriched in the constrained regions. The identification indexes are defined by sorting by *P*-value from the smallest to the largest. **(C)** Positions of identified motifs on the genome. The location of identified motifs was indicated by short black blocks in the ORFs. The vertical axis represents the credibility of the motif. **(D)** Motifs and related human RBPs. Motifs are colored in blue, and RBPs are colored in red. **(E)** GO terms enriched for motif-related human RBPs, including biological process, molecular function, and cellular component.

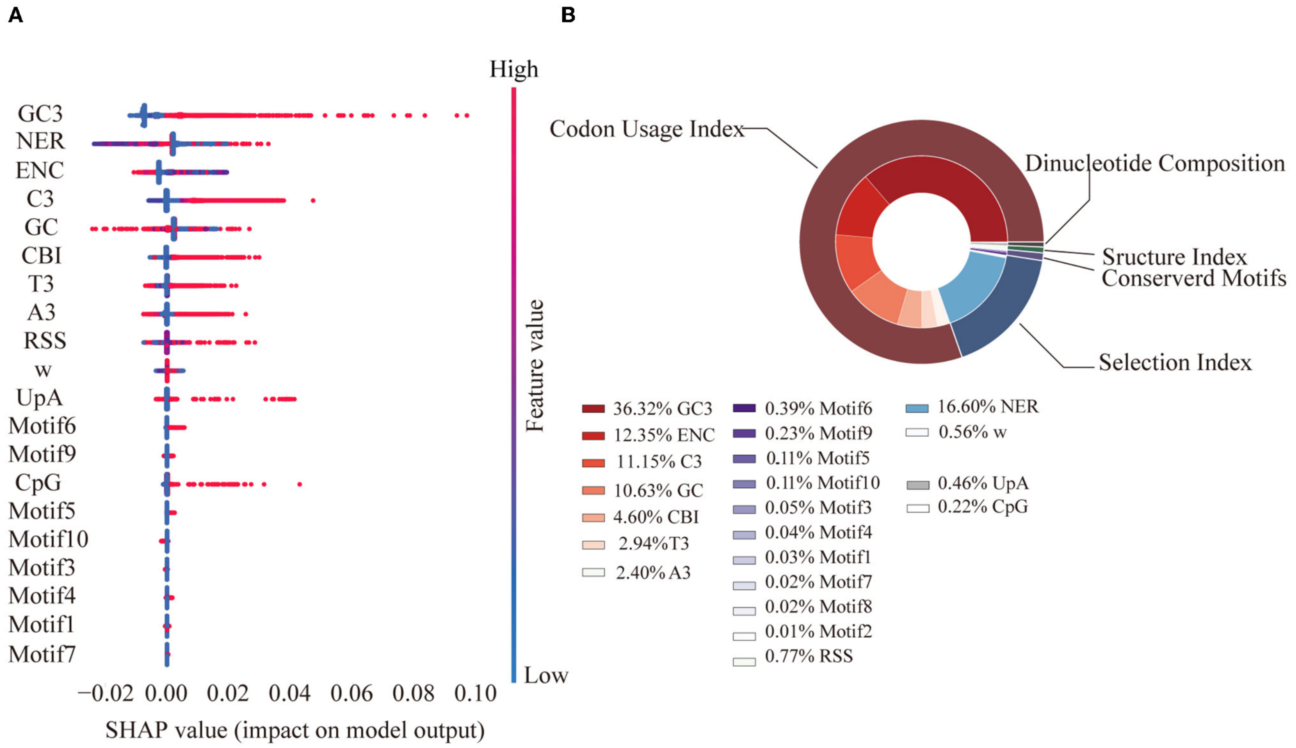


FIGURE 3
Feature SHAP value and contribution. **(A)** SHAP value for the top 20 features. Each point represents a sample. A SHAP value greater than 0 contributes to a higher SER, while a value less than 0 contributes to a lower SER. Feature value represents the value of each sample. **(B)** Feature importance pie chart. The outer ring represents the grouping, while the inner ring represents each specific feature. Percentage represents the proportion of the total interpretability.

the middle two groups, and the total number of positions in the two groups was found to be consistent and comparable. From a statistical viewpoint, the results of the four lineages that appeared first (Alpha, Beta, Gamma, and Delta) and were used to estimate the SER herein revealed that the characteristic mutations were significantly preferred in the partially constrained region than in the slightly free region (adjusted *chi-square*, $P = 0.036$) (Table 1). For the Omicron lineages, the sequences of which were not included in the SER estimation, characteristic mutations from the BA.2, BA.2.12.1, BA.4, and BA.5 sub-lineages showed a significant preference in the partially constrained regions, whereas the trend was not significant for BA.1 (adjusted *chi-square*, $P = 0.449$). For BA.2.75, a marginal *P*-value of 0.054 was obtained, indicating insufficient significance.

## Discussion

Viral synonymous changes are considered phenotypically silent, not functionally important, and frequently ignored; however, considering the continuing emergence of variants, it is necessary to speculate the significance of each type of mutation and its functional associations from the standpoint of synonymous substitutions, which are generally less studied. In this study, we found variations in the SERs across the SARS-CoV-2 genome. These variations can be partly explained by some factors, including the codon usage index, selection index, dinucleotide composition, structure index, and conserved motifs. Relevant motifs with extremely low SERs and potential functional constraints were identified in the constrained regions. Possible RBPs and their functions were also explored. The most important factor influencing the SER is the codon usage index. Fixed amino acid mutations are more likely to occur in partially constrained regions with potentially important functions and better adaptability. Our results indicated that the synonymous changes in the SARS-CoV-2 genome are not completely random and may be impacted by some fundamental functions and linked to the adaptation of future dominant variants.

Overall, the SERs in the SARS-CoV-2 genome vary across different regions. Their substitution rates ($0.4–1.0 \times 10^{-3}$ per site per year) (Figure 1A) are slightly lower than the traditionally observed substitution rates (approximately $10^{-4}$-$10^{-3}$ per site per year) based on the observed diversity (Boni et al., 2020; Chaw et al., 2020; Sharun et al., 2021; Singh and Yi, 2021), and the SER still follow the gamma distribution pattern (Kelly and Rice, 1996). The SERs were estimated using data from the first 2 years after SARS-CoV-2 infected the population. To achieve a certain level of adaptability after the virus has just infected the population, the virus will ensure a higher substitution rate than that in the equilibrium state, and this equilibrium state level may be closer to the estimated rate from synonymous sites. Statistically different SER distributions were also observed in several ORFs (Figure 1B) and different SER

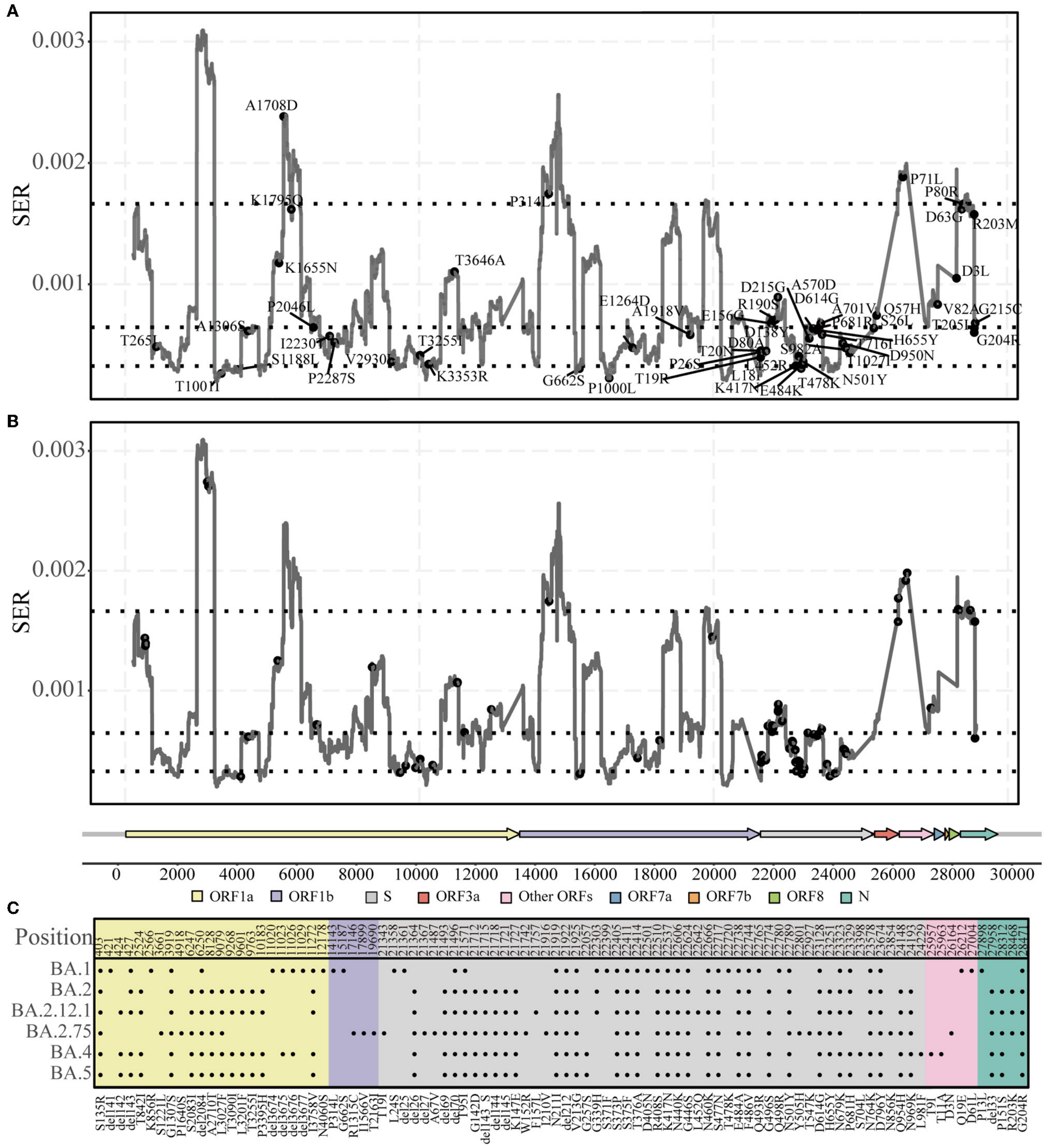


FIGURE 4
Characteristic mutations in five lineages. **(A)** Accumulated characteristic mutations in Alpha, Beta, Gamma, and Delta lineages. **(B, C)** Accumulated characteristic mutations in the Omicron lineage and their specific positions.

levels (Figure 1C) for positions. Discrepancies in the SERs between ORFs were also consistent with previous findings on *dS* estimation for other coronaviruses and SARS-CoV-2 (Singh and Yi, 2021; Wang H. et al., 2021).

In addition to the very high and very low SER values owing to the strong selection, we divided the middle 80% of the SERs into two groups. In contrast to the results obtained using traditional methods, where mutation events and entropy are considered, SER was found to be low in the S region in which diversity was previously thought to be high (compare Figure 1C and Supplementary Figure 3). The S protein is the most important surface protein in coronaviruses and is closely related to the virus

TABLE 1 Statistical test for positions of characteristic mutation accumulated in lineages.

| Region | Number of characteristic mutations | Number of not characteristic mutations | *P* value* |
|---|---|---|---|
| Alpha, Beta, Gamma and Delta | | | |
| Slightly free region | 15 | 3,307 | 0.036 |
| Partly constrained regions | 30 | 3,292 | |
| Omicron-BA.1 | | | |
| Slightly free region | 19 | 3,303 | 0.449 |
| Partly constrained regions | 25 | 3,297 | |
| Omicron-BA.2 | | | |
| Slightly free region | 13 | 3,309 | 0.005 |
| Partly constrained regions | 33 | 3,289 | |
| Omicron-BA.2.12.1 | | | |
| Slightly free region | 15 | 3,307 | 0.010 |
| Partly constrained regions | 34 | 3,288 | |
| Omicron-BA.2.75 | | | |
| Slightly free region | 16 | 3,306 | 0.054 |
| Partly constrained regions | 30 | 3,292 | |
| Omicron-BA.4 | | | |
| Slightly free region | 17 | 3,305 | 0.018 |
| Partly constrained regions | 35 | 3,287 | |
| Omicron-BA.5 | | | |
| Slightly free region | 13 | 3,309 | 0.003 |
| Partly constrained regions | 34 | 3,288 | |

*Adjusted chi-square test.

infectivity and pathogenesis (Andersen et al., 2020; Li Y. et al., 2021). The S protein has important evolutionary functions and functional constraints. However, owing to host switching and the long-term arms race with the host, this region experiences a certain degree of freedom, with a lot of changes occurring when it retains its original functions. Moreover, the higher diversity in the S region when counting mutation events or entropy may also be linked to the slightly deleterious mutations, which can later be removed by purifying selection. Furthermore, these measurements of diversity do not consider the rate of changes over time. However, from the SER viewpoint, the S protein region has important functions and certain adaptabilities, mostly in the partially constrained regions. All of these observations indicate that S protein changes impact the virus and could be related to adaptation.

Viruses have a simple structure, and they interact with appropriate hosts to cause infections. The viral genome plays a significant role when infecting a host (Ma-Lauer et al., 2012; Getts et al., 2013). The characteristics of conserved motifs from the constrained regions may indicate their functional importance during their interaction with a host. When matching the binding motif sites of human RBPs (Figure 2D), the identified motifs become associated with human RBPs, and some of the associated host RBPs have been identified and studied in previous coronavirus disease 2019-related studies. The knockdown of *YBX1*, which is associated with Motif3 and Motif6, reduces the viral RNA levels in both SARS-CoV-2 and Zika virus (Zhang et al., 2022). Together with *YBX1*, *ELAVL1*, which is found in viral RBP interactomes of SARS-CoV-2, is an IGF2BP1-related protein and a known mRNA stabilizer in humans, contributing to the stable translation of its target genes (Zhou and Pan, 2018). *SFPQ*, which interacts with the SARS-CoV-2 genome and promotes viral RNA amplification (Labeau et al., 2022), has been experimentally proven as a host factor required for the transcription of influenza virus; this can improve the transcription efficiency of viral mRNA polyadenylation (Landeras-Bueno et al., 2011). Furthermore, several *RBM* family proteins were involved in various steps of host RNA metabolism, including splicing, transportation, translation, and stability (Li Z. et al., 2021); moreover, the RBM family proteins were associated with the motifs identified in this study (Figure 2D). Functional annotation of these genes demonstrated their roles in RNA stabilization, binding single-stranded RNA, and translation regulation (Figure 2E). Our findings related to the conserved motifs from the constrained region and their potential functional importance provide a better understanding of the complete interaction landscape between the pathogen and host and may provide useful information for identifying novel drug or vaccine targets.

The features included in our model explained 72% of the SER variation. Among all the identified factors, sequence nucleotide and codon usage preferences were found to play a significant role

(Figure 3). Previous experiments in eukaryotes and prokaryotes have shown that codon usage bias is associated with gene expression and translation efficiency (Frumkin et al., 2018; Yang et al., 2019). The SARS-CoV-2 genome is AU-rich and has a clear preference for AU-rich codons over GC-rich codons; a similar trend has been observed in other coronavirus genomes, where UpA and CpG dinucleotides were strictly avoided. This may be attributable to the fact that viruses need to use host tRNA for translation and that the relative abundance of tRNAs in humans is inconsistent. Preference toward a certain nucleotide composition could improve viral translation efficiency in the host (Dilucca et al., 2020). Another explanation is that this bias may help viruses evade the innate immune response in humans (Roy et al., 2021). The significant number of synonymous transitions from C to U, which were reported in previous studies of the SARS-CoV-2 genome (De Maio et al., 2021; Morales et al., 2021) as well as observed in our study, was consistent with this phenomenon. The selection index substantially contributes to the variations in SERs (17.16%), with the single feature of the non-SER contributing the highest, indicating the importance of the contribution of selection pressure from the function requirement.

As new variants continue to emerge, previous studies have identified some characteristic mutations (including deletions) that are associated with viral transmissibility or infectivity (Bhattacharya et al., 2021; Kannan et al., 2021; Kumar et al., 2022; Papanikolaou et al., 2022). We found that the accumulated characteristic mutations mostly occurred in the partly constrained regions (Figure 4; Table 1); for example, the well-known P681H, Y505H, and E484K mutations occurred in the S region of many lineages. The location of the mutations in the partly constrained regions may play important roles; for example, they may alter the transmission rates and pathogenicity but simultaneously have the flexibility for tolerating mutations. Given that the Omicron genomes form a new monophyletic group (Kandeel et al., 2021), Omicron-related comparisons are more meaningful only when their sub-lineages are compared. For example, mutations are not significantly present in the partly constrained regions of Omicron BA.1; however, the opposite is observed for BA.2. Relevant studies have shown that BA.2 is more infectious than BA.1 (Elliott et al., 2022; Lyngse et al., 2022) and that the strains BA.2.75 and BA.2.12.1 exhibit the same phenomena as BA.2 (Table 1). These observations indicate that BA.1 may not be fully adapted as compared with the other lineages, owing to its sudden emergence. Mutations were indeed enriched in the partly constrained regions of BA.4 and BA.5. These strains are expected to become popular dominant strains and subsequently evolve into some new sub-lineages. One should pay careful attention to these sub-lineages, especially to BA.5, because the majority of their accumulated mutations have important functions. Thus, an estimate of the genomic SER can help quickly determine whether a mutation has significant impacts on circulation and could uniquely contribute toward rapid decision-making for preventing epidemics by compensating for the limitations of time-consuming laboratory tests.

Our study also has some limitations. (1) Our results are only based on the SARS-CoV-2 genome, and similar investigations in other viruses are warranted in the future. (2) The conserved motifs and their potential binding relationships with the host RBPs were mainly inferred through computational analyses, which require further experimental validation. (3) Some factors may not have been included in the SER variation analysis, which may have biased the understanding presented herein, and therefore, further investigation is warranted. (4) To identify important variants, other clues still need to be found and explored. Taken together, rather than ignoring synonymous mutations, one must pay further attention to them and explore the relationship between the synonymous mutations and other factors and the underlying mechanisms. All the relevant evidence gathered over time will ultimately help us to better prevent and control existing and future 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

XD conceived and designed the study and supervised the study. QS analyzed the data and drafted the manuscript. QS, JZen, and JZha collected the data. QS and JZha cleaned the data. QS, JZen, KT, HL, CZ, JZha, YX, JT, JZhe, SL, and LS commented on and revised the manuscript drafts. All authors read and approved the final report.

## Funding

This study was supported by the National Key Research and Development Projects of the Ministry of Science and Technology of China under grant 2021YFC2301300; the Shenzhen Science and Technology Program under grant KQTD20180411143323605, JSGG20200225152008136, and GXWD20201231165807008; the Guangdong Frontier and Key Tech Innovation Program under grants 2019B020228001, 2019B111103001, 2021A111112007, and 2022B1111020006; and the Natural Science Foundation of Guangdong Province, China, under grant 2021A1515011592.

## Acknowledgments

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing *via* the GISAID Initiative, on which this research is based.

## 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.2023.1136386/full#supplementary-material

## References

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9

Badua, C., Baldo, K. A. T., and Medina, P. M. B. (2021). Genomic and proteomic mutation landscapes of SARS-CoV-2. *J. Med. Virol.* 93, 1702–1721. doi: 10.1002/jmv.26548

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 37, 2834–2840. doi: 10.1093/bioinformatics/btab203

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–49. doi: 10.1093/nar/gkv416

Bhattacharya, M., Chatterjee, S., Sharma, A. R., Agoramoorthy, G., and Chakraborty, C. (2021). D614G mutation and SARS-CoV-2: impact on S-protein structure, function, infectivity, and immunity. *Appl. Microbiol. Biotechnol.* 105, 9035–9045. doi: 10.1007/s00253-021-11676-2

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* 5, 1408–1417. doi: 10.1038/s41564-020-0771-4

Callens, M., Pradier, L., Finnegan, M., Rose, C., and Bedhomme, S. (2021). Read between the lines: diversity of nontranslational selection pressures on local codon usage. *Genome Biol. Evol.* 13, evab097. doi: 10.1093/gbe/evab097

Chaw, S. M., Tai, J. H., Chen, S. L., Hsieh, C. H., Chang, S. Y., Yeh, S. H., et al. (2020). The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J. Biomed. Sci.* 27, 73. doi: 10.1186/s12929-020-00665-8

De Maio, N., Walker, C. R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., and Goldman, N. (2021). Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* 13, evab087. doi: 10.1093/gbe/evab087

de Oliveira, J. L., Morales, A. C., Hurst, L. D., Urrutia, A. O., Thompson, C. R. L., and Wolf, J. B. (2021). Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. *Mol. Biol. Evol.* 38, 3247–3266. doi: 10.1093/molbev/msab099

Dilucca, M., Forcelloni, S., Georgakilas, A. G., Giansanti, A., and Pavlopoulou, A. (2020). Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses* 12, 1–21. doi: 10.3390/v12050498

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276. doi: 10.1038/nrg2323

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* 1, 33–46. doi: 10.1002/gch2.1018

Elliott, P., Eales, O., Bodinier, B., Tang, D., Wang, H., Jonnerby, J., et al. (2022). Post-peak dynamics of a national Omicron SARS-CoV-2 epidemic during January 2022. *medRxiv* 2022.2002.2003.22270365. doi: 10.1101/2022.02.03.22270365

Flynn, R. A., Belk, J. A., Qi, Y., Yasumoto, Y., Wei, J., Alfajaro, M. M., et al. (2021). Discovery and functional interrogation of SARS-CoV-2 RNA-host protein interactions. *Cell.* 184, 2394–2411.e2316. doi: 10.1016/j.cell.2021.03.012

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., and Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc. Natl. Acad. Sci. USA* 115, E4940–e4949. doi: 10.1073/pnas.1719375115

Getts, D. R., Chastain, E. M., Terry, R. L., and Miller, S. D. (2013). Virus infection, antiviral immunity, and autoimmunity. *Immunol. Rev.* 255, 197–209. doi: 10.1111/imr.12091

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8, R24. doi: 10.1186/gb-2007-8-2-r24

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. doi: 10.1093/bioinformatics/bty407

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*.

Ho, S. Y., and Duchene, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* 23, 5947–5965. doi: 10.1111/mec.12953

Kandeel, M., Mohamed, M. E. M., Abd El-Lateef, H. M., Venugopala, K. N., and El-Beltagi, H. S. (2021). Omicron variant genome evolution and phylogenetics. *J. Med. Virol.* 94, 1627–1632. doi: 10.1002/jmv.27515

Kannan, S., Shaik Syed Ali, P., and Sheeza, A. (2021). Omicron (B.1.1.529)—variant of concern—molecular profile and epidemiology: a mini review. *Eur. Rev. Med. Pharmacol. Sci.* 25, 8019–8022. doi: 10.26355/eurrev_202112_27653

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

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36, D202–205. doi: 10.1093/nar/gkm998

Kelly, C., and Rice, J. (1996). Modeling nucleotide evolution: a heterogeneous rate analysis. *Math. Biosci.* 133, 85–109. doi: 10.1016/0025-5564(95)00083-6

Khan, M. T., Irfan, M., Ahsan, H., Ahmed, A., Kaushik, A. C., Khan, A. S., et al. (2021). Structures of SARS-CoV-2 RNA-binding proteins and therapeutic targets. *Intervirology* 64, 55–68. doi: 10.1159/000513686

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., et al. (2021). GISAID's role in pandemic response. *China CDC Wkly* 3, 1049–1051. doi: 10.46234/ccdcw2021.255

Kim, G., Shin, H. M., Kim, H. R., and Kim, Y. (2022). Effects of host and pathogenicity on mutation rates in avian influenza A viruses. *Virus Evol.* 8, doi: 10.1093/ve/veac013

Kumar, S., Thambiraja, T. S., Karuppanan, K., and Subramaniam, G. (2022). Omicron and Delta variant of SARS-CoV-2: a comparative computational study of spike protein. *J. Med. Virol.* 94, 1641–1649. doi: 10.1002/jmv.27526

Labeau, A., Fery-Simonian, L., Lefevre-Utile, A., Pourcelot, M., Bonnet-Madin, L., Soumelis, V., et al. (2022). Characterization and functional interrogation of the SARS-CoV-2 RNA interactome. *Cell Rep.* 39, 110744. doi: 10.1016/j.celrep.2022.110744

Landeras-Bueno, S., Jorba, N., Pérez-Cidoncha, M., and Ortín, J. (2011). The splicing factor proline-glutamine rich (SFPQ/PSF) is involved in influenza virus transcription. *PLoS Pathog.* 7, e1002397. doi: 10.1371/journal.ppat.1002397

Lauring, A. S., and Hodcroft, E. B. (2021). Genetic variants of SARS-CoV-2-What do they mean? *JAMA* 325, 529–531. doi: 10.1001/jama.2020.27124

Li, Y., Ma, M. L., Lei, Q., Wang, F., Hong, W., Lai, D. Y., et al. (2021). Linear epitope landscape of the SARS-CoV-2 Spike protein constructed from 1,051 COVID-19 patients. *Cell Rep.* 34, 108915. doi: 10.1016/j.celrep.2021.108915

Li, Z., Guo, Q., Zhang, J., Fu, Z., Wang, Y., Wang, T., et al. (2021). The RNA-binding motif protein family in cancer: friend or foe? *Front. Oncol.* 11, 757135. doi: 10.3389/fonc.2021.757135

Lu, S., Ye, Q., Singh, D., Cao, Y., Diedrich, J. K., Yates, J. R. 3rd, et al. (2021). The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.* 12, 502. doi: 10.1038/s41467-020-20768-y

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 56–67. doi: 10.1038/s42256-019-0138-9

Lyngse, F. P., Kirkeby, C. T., Denwood, M., Christiansen, L. E., Mølbak, K., Møller, C. H., et al. (2022). Transmission of SARS-CoV-2 Omicron VOC subvariants BA.1 and BA.2: evidence from Danish households. *medRxiv* 2022.2001.2028.22270044. doi: 10.1101/2022.01.28.22270044

MacLean, O. A., Lytras, S., Weaver, S., Singer, J. B., Boni, M. F., Lemey, P., et al. (2021). Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol.* 19, e3001115. doi: 10.1371/journal.pbio.3001115

Ma-Lauer, Y., Lei, J., Hilgenfeld, R., and von Brunn, A. (2012). Virus-host interactomes—antiviral drug discovery. *Curr. Opin. Virol.* 2, 614–621. doi: 10.1016/j.coviro.2012.09.003

McGrath, C. (2021). Synonymous but not equal: a special section and virtual issue on phenotypic effects of synonymous mutations. *Genome Biol. Evol.* 13, doi: 10.1093/gbe/evab186

Meng, G. K. Q., and Liu, T.-Y. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems (NIPS 2017)* (Long Beach, CA, USA.)

Morales, A. C., Rice, A. M., Ho, A. T., Mordstein, C., Mühlhausen, S., Watson, S., et al. (2021). Causes and consequences of purifying selection on SARS-CoV-2. *Genome Biol. Evol.* 13, evab196. doi: 10.1093/gbe/evab196

Mordstein, C., Cano, L., Morales, A. C., Young, B., Ho, A. T., Rice, A. M., et al. (2021). Transcription, mRNA Export, and Immune Evasion Shape the Codon Usage of Viruses. *Genome Biol. Evol.* 13, doi: 10.1093/gbe/evab106

Motozono, C., Toyoda, M., Zahradnik, J., Saito, A., Nasser, H., Tan, T. S., et al. (2021). SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* 29, 1124–1136.e1111. doi: 10.1016/j.chom.2021.06.006

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., et al. (2015). Gene-wide identification of episodic selection. *Mol. Biol. Evol.* 32, 1365–1371. doi: 10.1093/molbev/msv035

Oliver, J. L., Bernaola-Galván, P., Perfectti, F., Gómez-Martín, C., Verd,ú, M., and Moya, A. (2021). Accelerated decline of genome heterogeneity in the SARS-CoV-2 coronavirus. *bioRxiv* [preprint] 2021-11 doi: 10.1101/2021.11.06.467547

Papanikolaou, V., Chrysovergis, A., Ragos, V., Tsiambas, E., Katsinis, S., Manoli, A., et al. (2022). From delta to Omicron: S1-RBD/S2 mutation/deletion equilibrium in SARS-CoV-2 defined variants. *Gene* 814, 146134. doi: 10.1016/j.gene.2021.146134

Peden, J. F. (2000). *Analysis of Codon Usage.* CiteSeerX.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Pintó, R. M., and Bosch, A. (2021). The codon usage code for cotranslational folding of viral capsids. *Genome Biol. Evol.* 13, doi: 10.1093/gbe/evab089

Rahman, S., Kosakovsky Pond, S. L., Webb, A., and Hey, J. (2021). Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. *Proc. Natl. Acad. Sci. USA* 118:e2023575118. doi: 10.1073/pnas.2023575118

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177. doi: 10.1038/nature12311

Resch, A. M., Carmel, L., Mariño-Ramírez, L., Ogurtsov, A. Y., Shabalina, S. A., Rogozin, I. B., et al. (2007). Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* 24, 1821–1831. doi: 10.1093/molbev/msm100

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.* 11, 129. doi: 10.1186/1471-2105-11-129

Roy, A., Guo, F., Singh, B., Gupta, S., Paul, K., Chen, X., et al. (2021). Base composition and host adaptation of the SARS-CoV-2: insight from the codon usage perspective. *Front. Microbiol.* 12, 548275. doi: 10.3389/fmicb.2021.548275

Schmidt, N., Lareau, C. A., Keshishian, H., Ganskih, S., Schneider, C., Hennig, T., et al. (2021). The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat. Microbiol.* 6, 339–353. doi: 10.1038/s41564-020-00846-z

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

Sharun, K., Tiwari, R., Dhama, K., Emran, T. B., Rabaan, A. A., and Al Mutair, A. (2021). Emerging SARS-CoV-2 variants: impact on vaccine efficacy and neutralizing antibodies. *Hum. Vaccin. Immunother.* 17, 3491–3494. doi: 10.1080/21645515.2021.1923350

Shen, X., Song, S., Li, C., and Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*. 606, 725–731. doi: 10.1038/s41586-022-04823-w

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data—from vision to reality. *Euro Surveill.* 22, 30494. doi: 10.2807/1560-7917.Es.2017.22.13.30494

Singh, D., and Yi, S. V. (2021). On the origin and evolution of SARS-CoV-2. *Exp. Mol. Med.* 53, 537–547. doi: 10.1038/s12276-021-00604-z

Tao, K., Tzou, P. L., Nouhin, J., Gupta, R. K., de Oliveira, T., Kosakovsky Pond, S. L., et al. (2021). The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* 22, 757–773. doi: 10.1038/s41576-021-00408-x

Tzeng, Y. H., Pan, R., and Li, W. H. (2004). Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 21, 2290–2298. doi: 10.1093/molbev/msh242

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinform.* 8, 77–80. doi: 10.1016/s1672-0229(10)60008-3

Wang, H., Pipes, L., and Nielsen, R. (2021). Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evol.* 7, doi: 10.1093/ve/veaa098

Wang, Y., Zeng, J., Zhang, C., Chen, C., Qiu, Z., Pang, J., et al. (2021). New framework for recombination and adaptive evolution analysis with application to the novel coronavirus SARS-CoV-2. *Brief. Bioinform.* 22, doi: 10.1093/bib/bbab107

Wilson, D. J., Crook, D. W., Peto, T. E. A., Walker, A. S., Hoosdally, S. J., Gibertoni Cruz, A. L., et al. (2020). GenomegaMap: within-species genome-wide dN/dS estimation from over 10,000 genomes. *Mol. Biol. Evol.* 37, 2450–2460. doi: 10.1093/molbev/msaa069

Yang, Q., Yu, C. H., Zhao, F., Dang, Y., Wu, C., Xie, P., et al. (2019). eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic Acids Res.* 47, 9243–9258. doi: 10.1093/nar/gkz710

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, C., Wang, Y., Chen, C., Long, H., Bai, J., Zeng, J., et al. (2020). A mutation network method for transmission analysis of human influenza H3N2. *Viruses* 12, 1125. doi: 10.3390/v12101125

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., et al. (2020). SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* 11, 6013. doi: 10.1038/s41467-020-19808-4

Zhang, S., Huang, W., Ren, L., Ju, X., Gong, M., Rao, J., et al. (2022). Comparison of viral RNA-host protein interactomes across pathogenic RNA viruses informs rapid antiviral drug discovery for SARS-CoV-2. *Cell Res.* 32, 9–23. doi: 10.1038/s41422-021-00581-y

Zhou, K. I., and Pan, T. (2018). An additional class of m(6)A readers. *Nat. Cell Biol.* 20, 230–232. doi: 10.1038/s41556-018-0046-y

# 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

Frontiers in
Microbiology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact