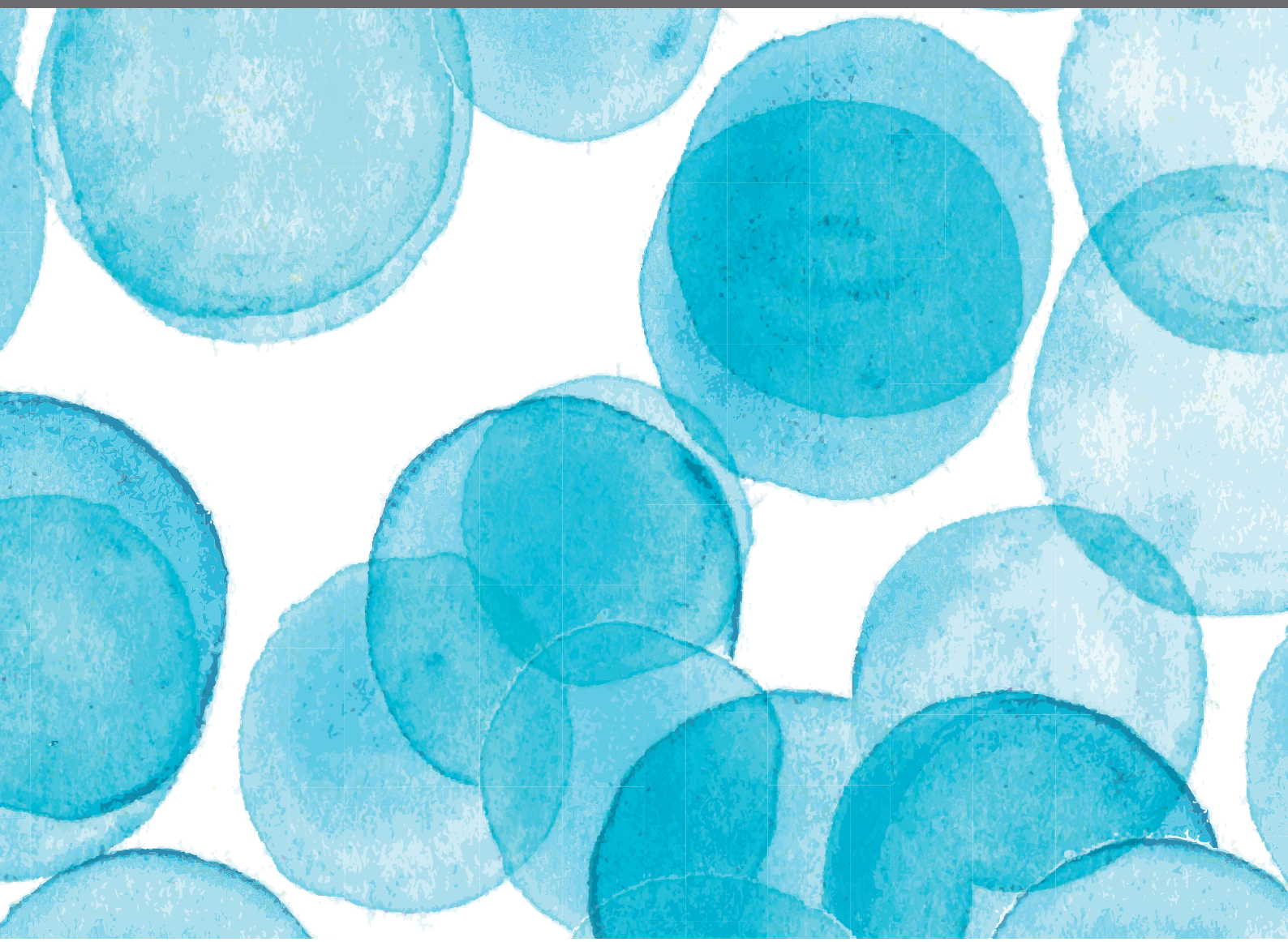
EVOLUTIONARY MECHANISMS OF INFECTIOUS DISEASES

EDITED BY: Yufeng Wang, Zhan Zhou and Jianying Gu
PUBLISHED IN: Frontiers in Microbiology,
Frontiers in Cellular and Infection Microbiology and
Frontiers in Genetics



frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

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

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

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

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

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

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

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

ISSN 1664-8714
ISBN 978-2-88966-985-1
DOI 10.3389/978-2-88966-985-1

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

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

Citation: Wang, Y., Zhou, Z., Gu, J., eds. (2021). Evolutionary Mechanisms of Infectious Diseases. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-985-1

Table of Contents

05 ***Editorial: Evolutionary Mechanisms of Infectious Diseases***
Jianying Gu, Zhan Zhou and Yufeng Wang

08 ***MicroGMT: A Mutation Tracker for SARS-CoV-2 and Other Microbial Genome Sequences***
Yue Xing, Xiao Li, Xiang Gao and Qunfeng Dong

12 ***Multilocus Sequence Typing of* Enterocytozoon bieneusi *Isolates From Various Mammal and Bird Species and Assessment of Population Structure and Substructure***
Xiaohua Liu, Yanchen Wu, Fengkun Yang, Baiyan Gong, Yanyan Jiang, Kexin Zhou, Jianping Cao, Weizhe Zhang, Aiqin Liu and Yujuan Shen

23 ***New* Plasmodium vivax *Genomes From the China-Myanmar Border***
Awtum M. Brashear, Adam C. Huckaby, Qi Fan, Luke J. Dillard, Yubing Hu, Yuling Li, Yan Zhao, Zenglei Wang, Yaming Cao, Jun Miao, Jennifer L. Guler and Liwang Cui

36 ***Comprehensive Genome Analysis of 6,000 USA SARS-CoV-2 Isolates Reveals Haplotype Signatures and Localized Transmission Patterns by State and by Country***
Lishuang Shen, Jennifer Dien Bard, Jaclyn A. Biegel, Alexander R. Judkins and Xiaowu Gai

40 ***Role of BgaA as a Pneumococcal Virulence Factor Elucidated by Molecular Evolutionary Analysis***
Masaya Yamaguchi, Moe Takemura, Kotaro Higashi, Kana Goto, Yujiro Hirose, Tomoko Sumitomo, Masanobu Nakata, Narikazu Uzawa and Shigetada Kawabata

51 ***Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2***
Sheng Liu, Jikui Shen, Shuyi Fang, Kailing Li, Juli Liu, Lei Yang, Chang-Deng Hu and Jun Wan

65 ***Comparative Evolutionary Patterns of* Burkholderia cenocepacia *and* B. multivorans *During Chronic Co-infection of a Cystic Fibrosis Patient Lung***
A. Amir Hassan, Sandra C. dos Santos, Vaughn S. Cooper and Isabel Sá-Correia

85 ***Characterization of the Complete Mitochondrial Genomes of Two Sibling Species of Parasitic Roundworms,* Haemonchus contortus *and* Teladorsagia circumcincta**
Nikola Palevich, Paul H. Maclean, Young-Jun Choi and Makedonka Mitreva

99 ***Molecular Pap Smear: Validation of HPV Genotype and Host Methylation Profiles of* ADCY8, CDH8, *and* ZNF582 *as a Predictor of Cervical Cytopathology***
Jane Shen-Gunther, Qingqing Xia, Winfred Stacey and Heisy B. Asusta

116 ***Differential Expression of Viral Transcripts From Single-Cell RNA Sequencing of Moderate and Severe COVID-19 Patients and Its Implications for Case Severity***
Teng Liu, Peilin Jia, Bingliang Fang and Zhongming Zhao

124 *G-Quadruplexes Are Present in Human Coronaviruses Including SARS-CoV-2*
Haoran Cui and Leiliang Zhang

132 *Host-Malaria Parasite Interactions and Impacts on Mutual Evolution*
Xin-zhuan Su, Cui Zhang and Deirdre A. Joy

152 *Epidemiological Characteristics and Formation Mechanisms of Multidrug-Resistant Hypervirulent* Klebsiella pneumoniae
Miran Tang, Xin Kong, Jingchen Hao and Jinbo Liu

162 *Identification of Uncommon* Cryptosporidium viatorum *(a Novel Subtype XVcA2G1c) and* Cryptosporidium andersoni *as Well as Common* Giardia duodenalis *Assemblages A and B in Humans in Myanmar*
Yanchen Wu, Baiyan Gong, Xiaohua Liu, Yanyan Jiang, Jianping Cao, Lan Yao, He Li, Aiqin Liu and Yujuan Shen

173 *Comparative Genomic Analyses Reveal a Specific Mutation Pattern Between Human Coronavirus SARS-CoV-2 and Bat-CoV RaTG13*
Longxian Lv, Gaolei Li, Jinhui Chen, Xinle Liang and Yudong Li

180 *Genomic Epidemiology of Antimalarial Drug Resistance in* Plasmodium falciparum *in Southern China*
Fang Huang, Christopher G. Jacob, Shannon Takala-Harrison, Matthew Adams, Heng-Lin Yang, Hui Liu, Zhi-Gui Xia, Shui-Sen Zhou, Lin-Hua Tang and Christopher V. Plowe

190 *Sensitive Immunoassay Detection of* Plasmodium *Lactate Dehydrogenase by Inductively Coupled Plasma Mass Spectrometry*
Jianbing Mu, Lee L. Yu and Thomas E. Wellems

197 *An Evolutionary Arms Race Between* Burkholderia pseudomallei *and Host Immune System: What Do We Know?*
Chalita Chomkatekaew, Phumrapee Boonklang, Apiwat Sangphukieo and Claire Chewapreecha

214 *Transcriptomic Basis of Serum Resistance and Virulence Related Traits in XDR* P. aeruginosa *Evolved Under Antibiotic Pressure in a Morbidostat Device*
Mumina Javed, Benedikt Jentzsch, Maximilian Heinrich, Viola Ueltzhoeffer, Silke Peter, Ulrich Schoppmeier, Angel Angelov, Sandra Schwarz and Matthias Willmann

228 *Evolution of Drug-Resistant* Mycobacterium tuberculosis *Strains and Their Adaptation to the Human Lung Environment*
Anna Allué-Guardia, Juan I. García and Jordi B. Torrelles

249 *Predominance of Non-carbapenemase Producing Carbapenem-Resistant Enterobacterales in South Texas*
Cody A. Black, Wonhee So, Steven S. Dallas, Gerard Gawrys, Raymond Benavides, Samantha Aguilar, Chang-Jui Chen, James F. Shurko and Grace C. Lee

frontiers in Microbiology

EDITORIAL
published: 13 May 2021
doi: 10.3389/fmicb.2021.667561

Editorial: Evolutionary Mechanisms of Infectious Diseases

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

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

Keywords: evolution, infectious diseases, infection, pathogenicity, virulence

Editorial on the Research Topic

Evolutionary Mechanisms of Infectious Diseases

OPEN ACCESS

Edited by:
Ludmila Chistoserdova, University of Washington, United States

Reviewed by:
Marina G. Kalyuzhanaya, San Diego State University, United States

***Correspondence:**
Jianying Gu
jianying.gu@csi.cuny.edu
Zhan Zhou
zhanzhou@zju.edu.cn
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: *13 February 2021*
Accepted: *15 April 2021*
Published: *13 May 2021*

Citation:
Gu J, Zhou Z and Wang Y (2021) Editorial: Evolutionary Mechanisms of Infectious Diseases. Front. Microbiol. 12:667561. doi: 10.3389/fmicb.2021.667561

Infectious diseases have been and continue to be one of the major causes of mortality around the world, posing significant health, social, and economic burdens. According to the World Health Organization (WHO), three infectious diseases (lower respiratory infections, diarrheal diseases, and tuberculosis) are among the top 10 causes of death. Infectious pathogens continue to emerge and re-emerge, underscoring considerable epidemic challenges to public health (Fauci, 2005; Watkins, 2018).

Evolution plays an important role in infectious diseases (Antia et al., 2003; Chabas et al., 2018; Echaubard et al., 2018). Driven by constant arms race between microbial pathogens and their hosts, pathogens evolve mechanisms to evade host defense (Hilleman, 2004; Van Avondt et al., 2015; Bernard et al., 2018), develop drug resistance (Nathan and Cars, 2014; Laxminarayan et al., 2016; Haldar et al., 2018), adapt to host environment (Sperandio, 2018; Beekman and Ene, 2020), compete with host microbiota (Vonaesch et al., 2018; Tsolis and Baumler, 2020), evolve virulence (Berngruber et al., 2013; Cressler et al., 2016; Geoghegan and Holmes, 2018), and spread and transmit to new hosts (McCallum et al., 2001; Antonovics et al., 2017). A better understanding of the key evolutionary features of infectious diseases such as pathogenicity, infectiousness, and transmissibility could result in effective prevention and control strategies.

Evolutionary processes of infectious diseases reflect temporal and spatial dynamic host-pathogen relationships, which may be revealed in genome structures and organization, population dynamics, and host/vector physiological systems. Recent efforts that bring together high-throughput omics technologies and big data analytical approaches have contributed to new insights of these processes (Cowell and Winzeler, 2019; Khan et al., 2019; Ball et al., 2020). The goal of this Research Topic is to provide a forum for sharing ideas, tools, and results among researchers from fields of evolutionary biology, infectious diseases, microbiology, genomics, and epidemiology. The Topic is mainly focused on recent progress in using multidisciplinary approaches to the studies of the evolutionary mechanisms of pathogenesis, virulence, host immunity, population dynamics, and epidemiology.

We sincerely thank all researchers who have contributed to our Research Topic. This collection of 21 articles is divided into four sections. The first section includes six articles focusing on Coronavirus disease 2019 (COVID-19) and its causative agent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of February 2021, over 104 million COVID-19 cases have been confirmed worldwide, including over 2.2 million deaths (https://covid19.who.int/) since

the emergence of the pandemic in December 2019 (Sohrabi et al., 2020; Yang et al., 2020). The six articles in this section offer new insight into the evolutionary and structural features of SARS-CoV-2: (1) Xing et al. develop a freely available program, MicroGMT, to identify and characterize mutations in microbial genomes, with the default setting optimized for SARS-CoV2. (2) Shen et al. present a comprehensive genomic epidemiology study that reveals a haplotype pattern of geographical specificity at the city-, state-, and country-levels, supporting the effectiveness of travel restriction in preventing widespread transmission. (3) Lv et al. describe a comparative genomic analysis between human SARS-CoV-2 and its close relative Bat-CoV RaTG13, and identify mutation patterns indicative of stronger purifying selection occurred in SARS-CoV-2. (4) The article by Liu S. et al. extends this topic. Distinct genomic single nucleotide variation signatures are identified in more than 30,000 SARS-CoV-2 genomes, which may have functional consequences due to the viral genetic instability (Liu S. et al.). (5) Readers interested in SARS-CoV-2 pathogenesis will appreciate a timely report by Liu T. et al. on the features and evolutionary difference of viral gene expression and fusion events in the SARS-CoV-2 infected cells from the patients with moderate and severe COVID-19. (6) Cui and Zhang report the presence of G-Quadruplexes in human coronaviruses including SARS-CoV-2, suggesting targeting G-Quadruplexes as a potential avenue to COVID-19 therapeutics.

The second section is devoted to the evolutionary mechanisms of bacteria. We are glad to present three review articles: Tuberculosis, one of the oldest known human infectious diseases, remains one of the major causes of mortality globally. Allué-Guardia et al. present a timely review to address the interactions of *Mycobacterium tuberculosis* with the lung environment and how these interactions may drive phenotypes of *M. tuberculosis* with a particular emphasis on drug resistant *M. tuberculosis*. Chomkatekaew et al. present the evidences and their perspective on the genetic and genomic diversity of *Burkholderia pseudomallei*, the cause of a lethal tropical disease melioidosis, and shed light on the evolutionary arms race between the bacterial pathogen and the host. The article by Tang et al. provides a succinct review regarding the pathogenesis and epidemiology of *Klebsiella pneumoniae*, the causative agent of a range of respiratory tract, urinary tract, and bloodstream infections. Two of the four original research articles are focused on evolutionary mechanisms of antibiotic resistance. The article by Black et al. highlights the increasing of non-carbapenemase producing Enterobacterales (NCPE) in south Texas. Their findings may have a direct impact on treatment regimens for patients. Javed et al. report a nice experimental evolution study of an extensively drug resistant (XDR) strain of *Pseudomonas aeruginosa*, an opportunistic pathogen leading to widespread infection and outbreak in hospitals. Their study demonstrates that the acquisition of antibiotic (colistin) resistance can affect the level of virulence (Javed et al.). Using a molecular evolutionary analysis approach, Yamaguchi et al. report the role of BgaA as a pneumococcal virulence factor in *Streptococcus pneumoniae*, a common cause of community-acquired pneumonia. Hassan et al. study and compare the genetic changes and adaptation that occur to two different *Burkholderia* species during co-infection of a cystic fibrosis patient's lungs over 4.4 years.

The third section is composed by seven articles on eukaryotic pathogens. Four articles are focused on malaria, a life-threatening mosquito-borne infectious disease. We highly recommend an excellent review article by Su et al. which addresses how the malaria parasites and their hosts interact, with a particular emphasis on the evolutionary impacts of recent human interventions such as drug and vaccine development. Three of the original articles address different important issues in malaria research: Brashear et al. present the draft of *de novo* genome assembly of four new *Plasmodium vivax* clinical samples collected in the China-Myanmar border area. The study provides new insight into genetic diversity and copy number variations of *P. vivax* multigene families (Brashear et al.). Huang et al. genotype the parasites from Southern China and present evidence for a genomic population with drug resistance. The study also demonstrates the utility of SNP microarrays for large-scale parasite molecular epidemiology (Huang et al.). Further on the application side, Mu et al. develop a novel, highly sensitive and accurate detection method for malaria parasite infection *via* integration of inductively coupled plasma mass spectrometry (ICP-MS) and Gold nanoparticles (AuNPs). Three articles in this section study evolutionary mechanisms of other eukaryotic pathogens. *Cryptosporidium* and *Giardia* cause intestinal illnesses in humans and animals. Wu et al. describe the first molecular epidemiological investigation of *Cryptosporidium* spp. and *G. duodenalis*, in humans in Myanmar, indicating the large potential of zoonotic transmission. The article by Palevich et al. describes a comprehensive analysis of the complete mitochondrial genomes of the New Zealand field strains of *Haemonchus contortus* and *Teladorsagia circumcincta*, two gastrointestinal nematodes. Using multilocus sequence typing, Liu X. et al. identify genetic diversity and clonal population structure in isolates of *Enterocytozoon bieneusi*, a unicellular microsporidian parasite closely related to fungi.

The last section of the collection includes an interesting piece of work by Shen-Gunther et al. on host-pathogen interactions. The article reports a large-scale validation study for the development of a panel of markers for the prediction of severity of cervical cancer (Shen-Gunther et al.). This panel, namely molecular pap smear, may have clinical applications since it captures pathogen and host specific signature related to pathogenicity and host susceptibility.

In summary, this collection of 21 articles covers a variety of topics in evolutionary mechanisms of infectious diseases, including the role of factors that influence pathogen virulence and host susceptibility, the role of genetic variation and population dynamics on pathogenesis, the role of medical interventions on drug resistance, and the role of disease control interventions on pathogen emergence and transmission. We hope that the Research Topic will be useful for a wide audience, particularly evolutionary biologists, microbiologists, infectious disease researchers and clinicians, genome scientists, systems biologists, graduate, and undergraduate students.

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.

REFERENCES

Antia, R., Regoes, R. R., Koella, J. C., and Bergstrom, C. T. (2003). The role of evolution in the emergence of infectious diseases. *Nature* 426, 658–661. doi: 10.1038/nature02104

Antonovics, J., Wilson, A. J., Forbes, M. R., Hauffe, H. C., Kallio, E. R., Leggett, H. C., et al. (2017). The evolution of transmission mode. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 372:20160083. doi: 10.1098/rstb.2016.0083

Ball, B., Langille, M., and Geddes-McAlister, J. (2020). Fun(gi)omics: advanced and diverse technologies to explore emerging fungal pathogens and define mechanisms of antifungal resistance. *mBio* 11. doi: 10.1128/mBio.01020-20

Beekman, C. N., and Ene, I. V. (2020). Short-term evolution strategies for host adaptation and drug escape in human fungal pathogens. *PLoS Pathog.* 16:e1008519. doi: 10.1371/journal.ppat.1008519

Bernard, Q., Smith, A. A., Yang, X., Koci, J., Foor, S. D., Cramer, S. D., et al. (2018). Plasticity in early immune evasion strategies of a bacterial pathogen. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3788–E3797. doi: 10.1073/pnas.1718595115

Berngruber, T. W., Froissart, R., Choisy, M., and Gandon, S. (2013). Evolution of virulence in emerging epidemics. *PLoS Pathog.* 9:e1003209. doi: 10.1371/journal.ppat.1003209

Chabas, H., Lion, S., Nicot, A., Meaden, S., van Houte, S., Moineau, S., et al. (2018). Evolutionary emergence of infectious diseases in heterogeneous host populations. *PLoS Biol.* 16:e2006738. doi: 10.1371/journal.pbio.2006738

Cowell, A. N., and Winzeler, E. A. (2019). Advances in omics-based methods to identify novel targets for malaria and other parasitic protozoan infections. *Genome Med.* 11:63. doi: 10.1186/s13073-019-0673-3

Cressler, C. E., McLeod, D. V., Rozins, C., Van Den Hoogen, J., and Day, T. (2016). The adaptive evolution of virulence: a review of theoretical predictions and empirical tests. *Parasitology* 143, 915–930. doi: 10.1017/S003118201500092X

Echaubard, P., Rudge, J. W., and Lefevre, T. (2018). Evolutionary perspectives on human infectious diseases: challenges, advances, and promises. *Evol. Appl.* 11,383–393. doi: 10.1111/eva.12586

Fauci, A. S. (2005). Emerging and reemerging infectious diseases: the perpetual challenge. *Acad. Med.* 80, 1079–1085. doi: 10.1097/00001888-200512000-00002

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

Haldar, K., Bhattacharjee, S., and Safeukui, I. (2018). Drug resistance in plasmodium. *Nat. Rev. Microbiol.* 16, 156–170. doi: 10.1038/nrmicro.2017.161

Hilleman, M. R. (2004). Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl. 2), 14560–14566. doi: 10.1073/pnas.0404758101

Khan, M. M., Ernst, O., Manes, N. P., Oyler, B. L., Fraser, I. D. C., Goodlett, D. R., et al. (2019). Multi-omics strategies uncover host-pathogen interactions. *ACS Infect. Dis.* 5, 493–505. doi: 10.1021/acsinfecdis.9b00080

Laxminarayan, R., Sridhar, D., Blaser, M., Wang, M., and Woolhouse, M. (2016). Achieving global targets for antimicrobial resistance. *Science* 353, 874–875. doi: 10.1126/science.aaf9286

McCallum, H., Barlow, N., and Hone, J. (2001). How should pathogen transmission be modelled? *Trends Ecol. Evol.* 16, 295–300. doi: 10.1016/S0169-5347(01)02144-9

Nathan, C., and Cars, O. (2014). Antibiotic resistance–problems, progress, and prospects. *N. Engl. J. Med.* 371, 1761–1763. doi: 10.1056/NEJMp1408040

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., et al. (2020). World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* 76, 71–76. doi: 10.1016/j.ijsu.2020.02.034

Sperandio, V. (2018). Pathogens' adaptation to the human host. *Proc. Natl. Acad. Sci. U.S.A.* 115, 9342–9343. doi: 10.1073/pnas.1813379115

Tsolis, R. M., and Baumler, A. J. (2020). Gastrointestinal host-pathogen interaction in the age of microbiome research. *Curr. Opin. Microbiol.* 53, 78–89. doi: 10.1016/j.mib.2020.03.002

Van Avondt, K., van Sorge, N. M., and Meyaard, L. (2015). Bacterial immune evasion through manipulation of host inhibitory immune signaling. *PLoS Pathog.* 11:e1004644. doi: 10.1371/journal.ppat.1004644

Vonaesch, P., Anderson, M., and Sansonetti, P. J. (2018). Pathogens, microbiome and the host: emergence of the ecological Koch's postulates. *FEMS Microbiol. Rev.* 42, 273–292. doi: 10.1093/femsre/fuy003

Watkins, K. (2018). Emerging infectious diseases: a review. *Curr. Emerg. Hosp. Med. Rep.* 6, 86–93. doi: 10.1007/s40138-018-0162-9

Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* 8, 475–481. doi: 10.1016/S2213-2600(20)30079-5

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.

Copyright © 2021 Gu, Zhou and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

TECHNOLOGY AND CODE
published: 25 June 2020
doi: 10.3389/fmicb.2020.01502

MicroGMT: A Mutation Tracker for SARS-CoV-2 and Other Microbial Genome Sequences

*Yue Xing[1], Xiao Li[2], Xiang Gao[3] and Qunfeng Dong[3,4]**

[1] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, United States, [2] Department of Molecular and Cellular Medicine, Texas A&M University, College Station, TX, United States, [3] Department of Medicine, Stritch School of Medicine, Loyola University Chicago, Maywood, IL, United States, [4] Center for Biomedical Informatics, Stritch School of Medicine, Loyola University Chicago, Maywood, IL, United States

With the continued spread of SARS-CoV-2 virus around the world, researchers often need to quickly identify novel mutations in newly sequenced SARS-CoV-2 genomes for studying the molecular evolution and epidemiology of the virus. We have developed a Python package, MicroGMT, which takes either raw sequence reads or assembled genome sequences as input and compares against database sequences to identify and characterize indels and point mutations. Although our default setting is optimized for SARS-CoV-2 virus, the package can be also applied to any other microbial genomes. The software is freely available at Github URL https://github.com/qunfengdong/MicroGMT.

Keywords: COVID-19, SARS-CoV-2, molecular evolution, epidemiology, bioinformatics, mutation, microbial, virus

OPEN ACCESS

Edited by:
Zhan Zhou,
Zhejiang University, China

Reviewed by:
Li Chen,
Indiana University, United States
Chen Chen,
Capital Medical University, China

***Correspondence:**
Qunfeng Dong
qdong@luc.edu

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

Received: *27 May 2020*
Accepted: *10 June 2020*
Published: *25 June 2020*

Citation:
Xing Y, Li X, Gao X and Dong Q (2020) MicroGMT: A Mutation Tracker for SARS-CoV-2 and Other Microbial Genome Sequences. Front. Microbiol. 11:1502. doi: 10.3389/fmicb.2020.01502

INTRODUCTION

Since the first submission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome sequence to GISAID on January 10th, 2020, there have already accumulated 34786 sequences at GISAID by May 31st, 2020 (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). The genome sequences of SARS-CoV-2 will continue accumulating since researchers heavily rely on DNA sequencing to study the epidemiology and molecular evolution of the virus (Lanciotti et al., 1994; Chinese SARS Molecular Epidemiology Consortium, 2004; Faye et al., 2014). SARS-CoV-2 is an RNA virus with limited proofreading capability of correcting replication errors (Domingo and Holland, 1997; Domingo et al., 2008; Duffy et al., 2008; Denison et al., 2011), thus it has been evolving continuously with new mutations (Bar-On et al., 2020; Koyama et al., 2020; Pachetti et al., 2020), which inevitably face selective pressures imposed by the host. Although most mutations are expected to be selectively neural, it is important to monitor if SARS-CoV-2 will eventually evolve to be a stronger or weaker infectious agent as time goes on. Therefore, it is vital to track mutations from newly sequenced SARS-CoV-2 genome.

Although whole genome sequencing of SARS-CoV-2 can be performed rapidly these days, bioinformatics tools are required to keep pace with the rapid sequencing capacity for tracking the mutations from newly sequenced genomes. To the best of our knowledge, there currently only exist two bioinformatics tools publicly available for detecting SARS-CoV2 mutations. The first is a web-based platform recently released by the Los Alamos National Laboratory (Korber et al., 2020). However, the tool is only limited for characterizing known mutations for the SARS-CoV-2 in genic regions. Users cannot submit their own sequences (e.g., newly sequenced viral genome isolated from COVID-19 patients) to compare with existing mutations in the database and identify new mutations. The second, named CoV-Glue, is also a web-based platform[1]. At the time of our

[1]http://cov-glue.cvr.gla.ac.uk/

manuscript preparation, CoV-Glue is still undergoing development and testing before its official release according to its web site. The main limitation with the current version of CoV-Glue is that it cannot detect mutations among multiple input sequences provided by users (only mutations between input and database sequences can be identified). In addition, both the above web-based tools can only analyze SARS-CoV-2 sequences instead of other microbial species. Therefore, we have developed a command-line based Python package, Microbial Genomics Mutant Tracker (MicroGMT), for users to detect mutations from their own sequence data (either for SARS-CoV-2 or other microbial genomes).

METHODS

Implementation

Figure 1 depicts the major computational steps of MicroGMT for either any microbial species or SARA-CoV-2. The required inputs to MicroGMT are (1) a fasta-format reference genome sequence, (2) a Genbank file, GTF or GFF-format annotation file for the reference genome, and (3) the user's own query sequences either as fastq-format raw sequence reads or fasta-format assembled genome sequences. The default reference genome of SARA-CoV-2 and its genome annotation are already included in the package [i.e., SARA-CoV-2 isolate Wuhan-Hu-1 complete genome sequence with GenBank accession number NC_045512 (Wu et al., 2020)]. We chose to perform pairwise alignment instead of multiple sequence alignment for two reasons: (1) biologically, researchers are mainly interested in using the reference genome as the coordinates to refer to the mutation. MicroGMT infers any genomic difference among non-reference genomes based on their individual alignment to the reference genome (i.e., do they carry the same or different mutation in a given genomic position with respect to the reference genome); (2) computationally, we can easily scale up the pairwise alignment by only computing each of new additional sequences instead of re-computing the entire multiple sequence alignment. This is an important computational issue to consider as there are already 34786 SARS-CoV-2 sequences in the database as of May 31st, 2020 and we expect that the number of database sequences will continue to grow rapidly. For SARS-CoV-2, we have already performed all the pairwise alignment for all the existing database sequences against the reference sequence and the results are already stored in pre-built summary tables as part of our package. When users download all the latest SARS-CoV-2 genome sequences (e.g., from GISAID.org), our package provides a utility script to compare their downloaded version with the precomputed ones in our package and perform the pairwise alignment only with the new ones. The BAM-format alignment outputs produced by minimap2 (Li, 2018) are used as input for the subsequent identification of genomic variants with bcftools (Li, 2011). If necessary, users can set the prior probability for the identification of variants. For user-supplied raw sequence reads, MicroGMT aligns them to the reference genome by the standard procedure consisting of BWA (Li and Durbin, 2009), Samtools (Li et al., 2009), and Picard[2] which performs alignment, sorting, duplication removal, and indel realignment for both single-end or paired-end sequence reads. For the aligned reads, GATK (McKenna et al., 2010; Van der Auwera et al., 2013) is used by MicroGMT for variant calling. Users can choose the cut-off of mapping quality and base quality for variant calling. All the identified mutations are listed in vcf-format outputs, to be characterized by snpEff (Cingolani et al., 2012) at both the DNA and protein levels for each mutation.

Based on the above results, MicroGMT automatically generates mutation summary tables, in which each identified mutation is further characterized as follows: (1) where is the genomic location of the mutation (e.g., genic or intergenic regions), (2) what is the mutation in the genomic and protein sequence (e.g., synonymous or non-synonymous), and (3) is it a novel mutation (i.e., never been identified in other existing genome sequences), or which/how many other existing genome sequences carry it already. If the optional regional file is provided by the user, this information is also incorporated in the summary tables: the software searches the sequence/sample ID in the provided file to extract regional information for the respective vcf-format output, and adds regional information to the sequence/sample ID in the table header. Using the summary tables, users can easily look at specific genes of interest by filtering the "gene" column of the summary tables. Users can also restrict their query within a specific genomic region of interest. For example, they may ask if there is any mutation in their supplied sequences within the region of Spike gene by restricting their query for the genomic coordinates of the Spike gene. MicroGMT was built under Python 3 and can be used in Linux/Mac OS systems.

Validation

We simulated mutations in 1000 SARS-CoV-2 genome sequences and their corresponding raw sequence reads to validate the performance of MicroGMT. The complete genome sequence of SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank accession number NC_045512) was used as the reference genome for simulating mutations. The simuG program (Yue and Liti, 2019) was used to simulate random point mutations and indels to the reference genome sequence. The number of mutations in each simulated genome sequence was drawn from a Poisson distribution based on the average number of mutations per genome that we observed among real SARS-CoV-2 sequences in GISAID. Using those simulated genome sequences, their corresponding raw sequence reads were also simulated using ART-illumina (Huang et al., 2012) with the following parameters: sequencing platform of HiSeq2500, fold coverage of 50, paired-end read length of 150 bp, average fragment size of 300 bp with standard deviation of 20 bp, and base quality between 18 and 38 (**Supplementary File 1**). For evaluating the validation results, sensitivity is defined as the number of genomic mutations identified by MicroGMT divided by the number of true mutations introduced via simulation; specificity is defined as the number of genomic loci without mutations identified by MicroGMT divided by

[2]https://broadinstitute.github.io/picard/

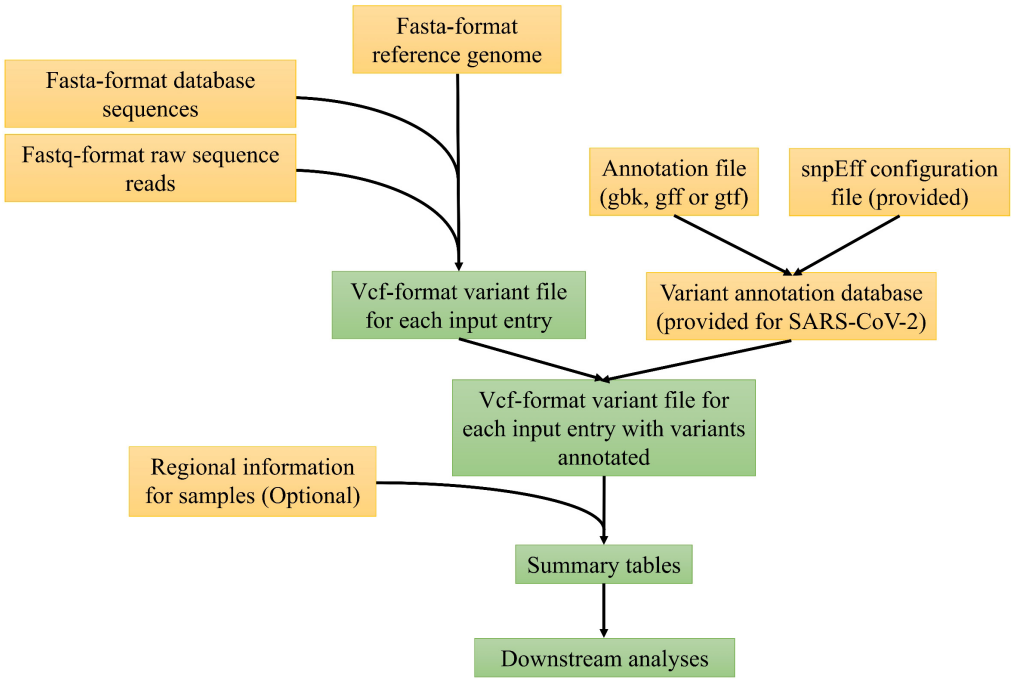


FIGURE 1 | The overview of MicroGMT workflow. See main text for details.

the number of genomic loci without any simulated mutations; accuracy is defined as the sum of genomic loci with and without mutations identified by MicroGMT divided by the total number of genomic loci with and without simulated mutations.

RESULTS

To validate the reliability of MicroGMT for mutation identification, 1000 artificial mutant strains of SARS-CoV-2 were simulated in which point mutations and/or indels were introduced randomly based on a Poisson distribution with the observed average number of mutations per genome in the real dataset of GISAID. By applying MicroGMT to this simulated test dataset, we obtained a specificity of 100%, a sensitivity of 100%, and an accuracy of 100% for both simulated whole genome sequences and raw sequence reads. In other words, MicroGMT is highly reliable in correctly identifying all mutations.

For real-world SARS-CoV-2 sequences, we would not know the exact genomic position of each potential mutation as in the simulated dataset. Therefore, we evaluated the reliability of MicroGMT through random sampling and manual verification. Specifically, we applied MicroGMT to ten randomly selected SARS-CoV-2 whole genome sequences, as well as their simulated paired-end raw sequence reads using ART-illumina (fold coverage of 10, read length of 150 bp, average fragment size of 300 bp with standard deviation of 20 bp) as test samples. Sensitivity and accuracy were both 100% for whole genome sequences, and were 98.4 and 100% for simulated raw sequence reads, respectively. The only point mutation that was not identified by MicroGMT for raw sequence reads was only 28 base pairs away from a stretch of unassigned bases (Ns), making it difficult to align the raw reads at that region (a problem with quality of the raw reads).

To illustrate the utility of MicroGMT, we applied it to identify all the mutations in a total of 34786 strains of SARS-CoV-2 in the GISAID database as of May 31st, 2020. In total, 15312 unique mutations were identified from 12618 genomic loci in SARS-CoV-2 reference genome (14749 mutations were located inside genes at 12245 loci). The running time of MicroGMT scales linearly with the number of input sequences. It took roughly 40 h to process all 34786 sequences. For a user with 100 input sequences, it would take less than 7 min for assembled genome sequences and less than 30 min for raw sequence reads with 50x depth of coverage.

DISCUSSION

MicroGMT can identify, annotate and summarize mutations in a large number of microbial genome sequences. Although it is optimized for SARS-CoV-2 sequences, it can be easily applied to other microbial sequences (potentially any haploid genomes). By tracking mutations among closely related strains, MicroGMT identifies where and what kinds of the mutations are at both DNA and protein levels.

MicroGMT is designed for tracking indels and SNPs among closely related strains instead of detecting large-scale complex genomic rearrangements and duplications. In addition, if users supply fastq-format raw sequence reads as input, the accuracy of mutation detection can be slightly affected by unmasked repetitive regions in the reference genome (results not shown) due to the difficult nature of aligning short sequence reads to the reference genome.

Overall, MicroGMT provides a fast, flexible, and reliable solution for researchers who need to track genomic mutations for SARS-CoV-2 virus or other microbial species.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets analyzed for this study is available at the GISAID website at https://www.gisaid.org/. The MicroGMT software is freely available at https://github.com/qunfengdong/MicroGMT.

AUTHOR CONTRIBUTIONS

YX implemented the software and drafted the manuscript. XL and YX performed the testing and data analysis. XG and QD conceived the project and design. All authors contributed to the article and approved the submitted version.

FUNDING

XG and QD were partially supported by NIH grant 5R01AI116706.

ACKNOWLEDGMENTS

The high-performance research computer Ada at Texas A&M University was used to test the MicroGMT software. We would like to thank Dr. Jiarui Li at the Beijing Ditan Hospital in China for independently testing MicroGMT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01502/full#supplementary-material

REFERENCES

Bar-On, Y. M., Flamholz, A., Phillips, R., and Milo, R. (2020). SARS-CoV-2 (COVID-19) by the numbers. *eLife* 9:e5309.

Chinese SARS Molecular Epidemiology Consortium (2004). Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 303, 1666–1669. doi: 10.1126/science.1092002

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

Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D., and Baric, R. S. (2011). Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* 8, 270–279. doi: 10.4161/rna.8.2.15013

Domingo, E., Grande-Pérez, A., and Martín, V. (2008). Future prospects for the treatment of rapidly evolving viral pathogens: insights from evolutionary biology. *Expert Opin. Biol. Ther.* 8, 1455–1460. doi: 10.1517/14712598.8.10.1455

Domingo, E., and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51, 151–178. doi: 10.1146/annurev.micro.51.1.151

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. Challenges* 1, 33–46. doi: 10.1002/gch2.1018

Faye, O., Freire, C. C., Iamarino, A., Faye, O., de Oliveira, J. V. C., Diallo, M., et al. (2014). Molecular evolution of Zika virus during its emergence in the 20th century. *PLoS Negl. Trop. Dis.* 8:e2636. doi: 10.1371/journal.pntd.0002636

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi: 10.1093/bioinformatics/btr708

Korber, B., Fischer, W., Gnanakaran, S. G., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioexiv [Preprint]* doi: 10.1101/2020.04.29.069054

Koyama, T., Platt, D., and Parida, L. (2020). *Variant Analysis of COVID-19 Genomes*. Yotktown Heights, NY: IBM TJ Watson Research Center.

Lanciotti, R. S., Lewis, J. G., Gubler, D. J., and Trent, D. W. (1994). Molecular evolution and epidemiology of dengue-3 viruses. *J. Gen. Virol.* 75, 65–75. doi: 10.1099/0022-1317-75-1-65

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

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

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

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18, 1–9.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* 22:30494.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1-11.10.33.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., et al. (2020). *Data from: Direct Submission of Severe Acute Respiratory Syndrome Coronavirus 2 isolate Wuhan-Hu-1, Complete Genome to RefSeq*. Shanghai: Fudan University.

Yue, J. X., and Liti, G. (2019). simuG: a general-purpose genome simulator. *Bioinformatics* 35, 4442–4444. doi: 10.1093/bioinformatics/btz424

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.

Copyright © 2020 Xing, Li, Gao and Dong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 June 2020
doi: 10.3389/fmicb.2020.01406

Multilocus Sequence Typing of *Enterocytozoon bieneusi* Isolates From Various Mammal and Bird Species and Assessment of Population Structure and Substructure

Xiaohua Liu[1†], Yanchen Wu[1†], Fengkun Yang[1], Baiyan Gong[1], Yanyan Jiang[2], Kexin Zhou[1], Jianping Cao[2], Weizhe Zhang[1], Aiqin Liu[1] and Yujuan Shen[2*]*

[1] Department of Parasitology, Harbin Medical University, Harbin, China, [2] National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Chinese Center for Tropical Diseases Research, WHO Collaborating Centre for Tropical Diseases, National Center for International Research on Tropical Diseases, Ministry of Science and Technology, Key Laboratory of Parasite and Vector Biology, MOH, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Longxian Zhang,
Henan Agricultural University, China
Guangneng Peng,
Sichuan Agricultural University, China

***Correspondence:**
Aiqin Liu
liuaiqin1128@126.com
Yujuan Shen
amyshyj12@163.com

[†] 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: 09 April 2020
Accepted: 29 May 2020
Published: 26 June 2020

Citation:
Liu X, Wu Y, Yang F, Gong B, Jiang Y, Zhou K, Cao J, Zhang W, Liu A and Shen Y (2020) Multilocus Sequence Typing of Enterocytozoon bieneusi Isolates From Various Mammal and Bird Species and Assessment of Population Structure and Substructure. Front. Microbiol. 11:1406. doi: 10.3389/fmicb.2020.01406

Enterocytozoon bieneusi is one of the most common intestinal pathogens in humans and animals. *E. bieneusi* has been confirmed to be complex microsporidian species. Approximately 500 ITS genotypes of *E. bieneusi* have been defined. With the establishment and application of multilocus sequencing typing and population genetic tools in *E. bieneusi*, the studies on these aspects have been carried out worldwide, but little information is available. To understand genetic variation of mini-/micro-satellites and the population structure and substructure of *E. bieneusi* in northeastern China, 305 *E. bieneusi* DNA specimens composed of 28 ITS genotypes were from 13 mammal species and five bird species in the investigated areas. They were characterized by nested-PCR amplification and sequencing at four mini-/micro-satellite loci (MS1, MS3, MS4, and MS7). At the MS1, MS3, MS4, and MS7 loci, 153 (50.16%), 131 (42.95%), 133 (43.61%), and 128 (41.97%) DNA specimens were amplified and sequenced successfully with 44, 17, 26, and 24 genotypes being identified, respectively. Multilocus genotypes (MLGs) showed a higher genetic diversity than ITS genotypes. 48 MLGs were produced out of 90 ITS-positive DNA specimens based on concatenated sequences of all the five genetic loci including ITS. Linkage disequilibrium (LD) and limited genetic recombination were observed by measuring LD using both multilocus sequences and allelic profile data, indicating an overall clonal population structure of *E. bieneusi* in the investigated areas. These data will aid in the longitudinal tracking of the attribution of source of infection/contamination and in elucidating transmission dynamics, and will provide valuable information for making efficient control strategies to intervene with and prevent occurrence of microsporidiosis caused by

E. bieneusi among animals and transmission of *E. bieneusi* from animals to humans in the investigated areas. Phylogenetic and network analyses identified three different subgroups, revealing the presence of host-shaped segregation and the absence of geographical segregation in *E. bieneusi* population. Meanwhile, the MLGs from zoonotic ITS genotypes were observed to be basically separated from the MLGs from host-adapted ones. Assessment of substructure will have a reference effect on understanding of zoonotic or interspecies transmission of *E. bieneusi* and evolution direction from zoonotic genotypes to host-adapted genotypes.

Keywords: *Enterocytozoon bieneusi*, multilocus sequence typing, population structure, population substructure, evolution

INTRODUCION

Enterocytozoon bieneusi is the most common microsporidian species (more than 90%) in reported cases of human microsporidiosis worldwide (Matos et al., 2012). Microsporidiosis caused by *E. bieneusi* is characterized by diarrhea, and the severity of diarrhea is closely related to the health status of the infected hosts. Self-limiting diarrhea usually appears in immunocompetent or healthy individuals while chronic or life-threatening diarrhea often occurs in immunocompromised/immunodeficient individuals, such as acquired immune deficiency syndrome (AIDS) patients, organ transplant recipients, cancer patients, travelers, children, and the elderly (Matos et al., 2012; Li and Xiao, 2019; Li et al., 2019a). In addition to humans, *E. bieneusi* has been detected in a variety of animal species, raising a concern of zoonotic transmission (Santín and Fayer, 2011; Wang et al., 2018b). The findings of the same genotypes of *E. bieneusi* in both humans and animals support the presumption of zoonotic potentials (Li et al., 2019b).

E. bieneusi has been confirmed to be a complex species with multiple genotypes. With application of genotyping tools and development of genetic markers related to *E. bieneusi*, a deep understanding has been reached about the host specificity and evolution of *E. bieneusi* (Li et al., 2019b). Based on sequence analysis of the internal transcribed spacer (ITS) region of the ribosomal RNA (rRNA) gene of numerous *E. bieneusi* isolates from humans and animals, to date, approximately 500 ITS genotypes of *E. bieneusi* have been identified, and phylogenetic analysis have revealed 11 different phylogenetic groups (groups 1–11) and an outlier with different degrees of host specificity (Li et al., 2019b). Currently, approximately 40 genotypes in group 1 have been identified in humans and multiple animal species, suggesting a large probability of cross-species transmission risks, therefore, this group is considered to be a zoonotic group (Shen et al., 2020). Some genotypes in group 2 (notably BEB4, BEB6, I, and J) have a broader host range than originally thought, increasing their importance for public health. The other genotypes in groups 3–11 seem to have limited or minimal effects on public health due to their more host-specific adaptation (Li et al., 2019a). However, the single ITS locus (243 bp in length) has limitations in representing the whole genome of *E. bieneusi* (∼6 Mb total length), for it does not uncover the subtle changes among *E. bieneusi* isolates closely genetically related to each other at the ITS locus (Widmer and Akiyoshi, 2010; Feng et al., 2011). Therefore, these observations need to be substantiated by sequence characterization of other genetic markers (Santín and Fayer, 2009). A multilocus sequencing tool for high-resolution typing of *E. bieneusi* has been established and widely used to characterize the population structures of *E. bieneusi* isolates based on length polymorphisms and single nucleotide polymorphisms (SNPs) of three microsatellites (MS1, MS3, and MS7) and one minisatellite (MS4) (Feng et al., 2011). To date, MLST analysis have been performed in at least 167 ITS genotypes of *E. bieneusi* in at least ten countries, with 23 from humans and 151 from animals (**Supplementary Table S1**). Some MLST data have been used to assess population structures of *E. bieneusi* isolates derived from different hosts: AIDS patients (Li et al., 2012, 2013), non-human primates (NHPs) (Karim et al., 2014), pigs (Wan et al., 2016; Wang et al., 2018a; Li et al., 2019; Zhang et al., 2020), foxes and raccoon dogs (Li et al., 2016b), captive giant pandas (Li et al., 2017), and dairy cattle (Wang et al., 2019). Meanwhile, some studies also explored the issues of *E. bieneusi* population genetics in relation to geographical origins and host origins (Li et al., 2013, 2016b, 2017, 2019; Wan et al., 2016; Wang et al., 2018a, 2019; Zhang et al., 2020).

In China, since the first identification of *E. bieneusi* in 2011 (Zhang et al., 2011), molecular epidemiological studies of *E. bieneusi* have been carried out in humans (prevalence: 0.2–22.5%) and animals (prevalence: 0–100%) in at least eight and 29 provinces, municipalities or autonomous regions, respectively, (Gong et al., 2019; Qiu et al., 2019). In northeastern China, *E. bieneusi* is prevalent in mammal and bird species, with the prevalence high up to 89.5% in asymptomatic pigs (Zhao et al., 2014b). However, only two studies reported multilocus genotypes (MLGs) of *E. bieneusi* isolates from two animal species (Wan et al., 2016; Zhang et al., 2016). To further understand the population structure and substructure of *E. bieneusi* in China, 305 *E. bieneusi* isolates, derived from 18 various animal species (mammals and birds) in northeastern China, were characterized by MLST analysis at four mini-/micro-satellite loci (MS1, MS3, MS4, and MS7). Meanwhile, linkage disequilibrium (LD) estimation and phylogenetic and network analyses were also conducted. These results will provide valuable information for making efficient control strategies to intervene with and prevent occurrence of microsporidiosis caused by *E. bieneusi* among animals and transmission of *E. bieneusi* from animals to humans,

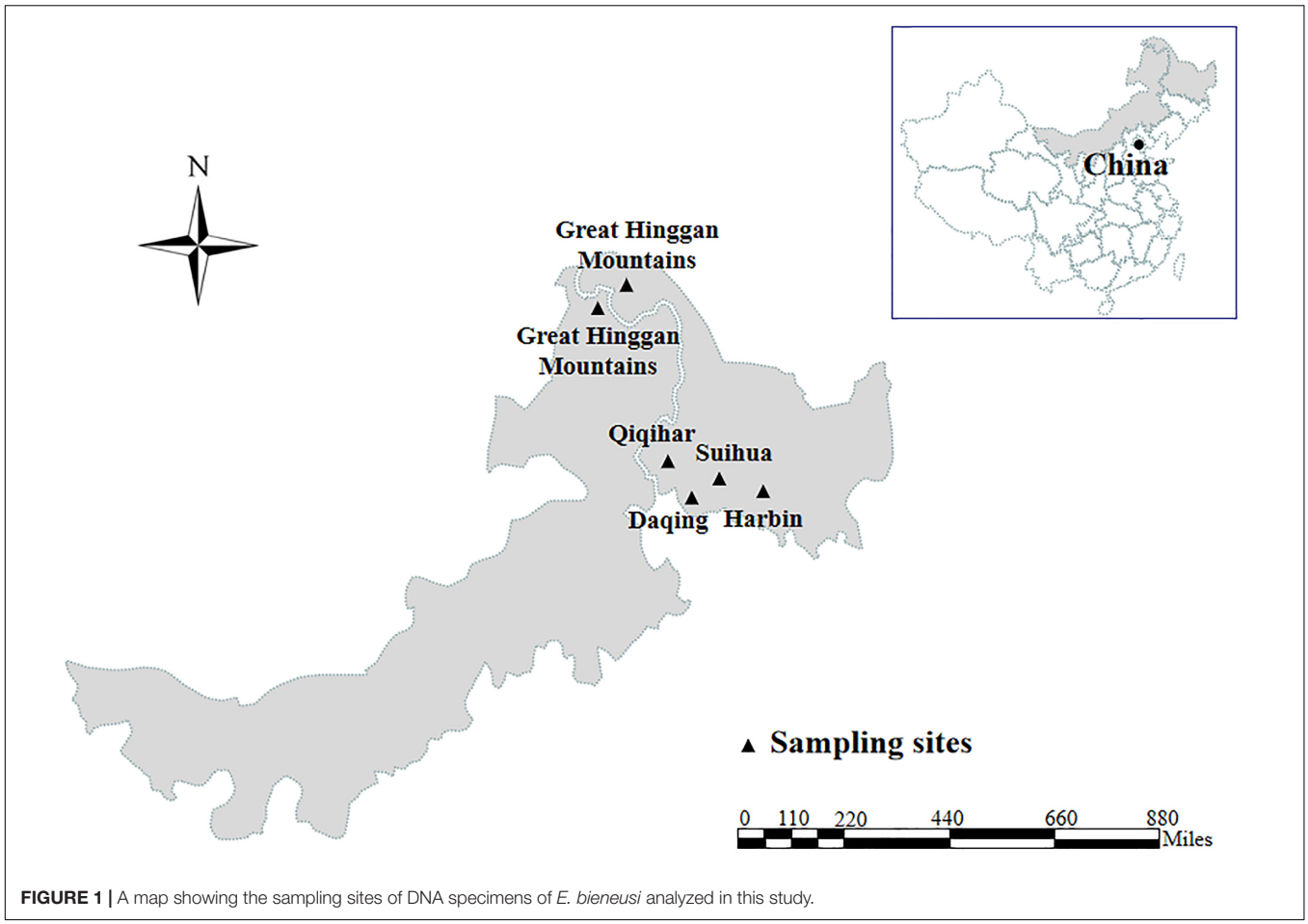


FIGURE 1 | A map showing the sampling sites of DNA specimens of *E. bieneusi* analyzed in this study.

and will improve our perception about the population structure and substructure of this pathogen.

MATERIALS AND METHODS

Ethics Statement

This study uses *E. bieneusi* isolates obtained from animals in northeastern China. The Research Ethics Committee and the Animal Ethical Committee of Harbin Medical University did not require the study to be reviewed or approved by an ethics committee because all the specimens analyzed in this study were DNA preparations.

E. bieneusi DNA Specimens

Enterocytozoon bieneusi genomic DNA was extracted from each of animal fecal specimens (252 from 13 mammal species and 53 from five bird species) distributing in four cities and Great Hinggan Mountains in northeastern China during the period of May 2012 to June 2019 using a QIAamp DNA Mini Stool Kit (Qiagen, Hilden, Germany) according to the manufacturer recommended procedures (**Figure 1**). ITS genotypes were identified by nested PCR amplification of an approximately 390 bp nucleotide fragment of the rRNA gene including 243 bp of the ITS region (Buckholt et al., 2002). All the DNA specimens were composed of 28 ITS genotypes belonging to group 1 (n = 20), group 2 (n = 7) and group 9 (n = 1), with some of them having been published in our previous studies (Zhao et al., 2014a,b, 2015a,b,c, 2016, 2017, 2018; Liu et al., 2015; Yang et al., 2016) (**Supplementary Table S2**). The ITS genotypes, numbers, host origins and phylogenetic groups of all isolates used in the present MLST analysis were showed in **Supplementary Table S3**.

TABLE 1 | Amplification efficiency and genetic polymorphism of *E. bieneusi* isolates (n = 305) at the MS1, MS3, MS4, and MS7 loci.

Locus	No. of positives (amplification efficiency)	No. of genotypes	Genetic polymorphism	
			Repeat (n)	No. of SNPs
MS1	153 (50.16%)	44	TGC (1); TAA (5–38); T(C)AC(T) (3–10)	14
MS3	131 (42.95%)	17	TA (4–17)	36
MS4[a]	133 (43.61%)	26	G(A)G(A)TA (1–2)	39
MS7	128 (41.97%)	24	TAA (3–11)	24

[a]*Insertions and deletions (INDELs) were observed at the MS4 locus.*

Multilocus Sequence Typing

Except for the ITS region of the rRNA gene, all the 305 *E. bieneusi* DNA specimens were also analyzed at the four MLST loci, including three microsatellites (MS1, MS3, and MS7) and one minisatellite (MS4). The primers and the cycle parameters for nested PCR amplifications used in the present study were designed previously by Feng et al. (2011) and the approximately expected fragment lengths of the secondary PCR products were 676 for MS1, 537 for MS3, 885 for MS4 and 471 for MS7 (Feng et al., 2011). TaKaRa Taq DNA polymerase (TaKaRa Bio Inc., Tokyo, Japan) was used in all the PCR reactions. All secondary PCR products were separated by electrophoresis in a 1.5% agarose gel and visualized on GelDoc^{TM} EZ Imager (Bio-Rad, United States) by staining the gel with GelStrain (TransGen Biotech., Beijing, China) before sequencing.

DNA Sequence Analysis

All secondary PCR products of expected size were directly sequenced with their respective PCR primers on an ABI PRISM^{TM} 3730 DNA Analyzer by Sinogeno-max Biotechnology Co., Ltd. (Beijing, China), using a BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA, United States). Two-directional sequencing method was performed to make sure of sequence accuracy. Nucleotide sequences obtained in the present study were compared to all *E. bieneusi* homologous sequences published in GenBank using BLAST searches[1]. They were then aligned and analyzed with each other and reference sequences downloaded from the GenBank database using the program Clustal X 1.83[2] to determine genotypes at the ITS locus and the genotypes at the MS1, MS3, MS4, and MS7 loci, respectively.

Analysis of Population Structure

The standardized index of association (I^S_A) proposed by Haboud and Hudson was used to assess the multilocus LD over the allelic profile data using software LIAN 3.5[3]. I^S_A equal to zero or with a negative value indicates randomly mating populations and alleles in linkage equilibrium (LE). In contrast, when an I^S_A value is greater than zero, there is the existence of non-panmictic population structure and LD in allelic profile data analysis. In addition, the population structure was also assessed by calculating the variance of pairwise differences (V_D) in another test and the 95% critical value (L) for V_D. When the V_D is less than L, the population is panmictic and is in LE; otherwise, the population is non-panmictic and a certain degree of LD exists (Haubold et al., 1998). Based on the fact of a very high degree of genetic polymorphism in the ITS region of the rRNA gene within *E. bieneusi*, to assess the influence of the ITS gene on the result of LD analysis, multilocus LD analyses were assessed over the allelic data including and excluding the ITS gene. Gene diversity (Hd), intragenic LD and recombination events (Rms) were also calculated from the concatenated sequences including and excluding the ITS gene without consideration of short insertions and deletions (INDELs) by using DnaSP 5.10.01[4]. The population structure of *E. bieneusi* isolates was assessed by measuring the intragenic LD, I^S_A, and Rms.

Analysis of Population Substructure

Phylogenetic relationship of MLGs of *E. bieneusi* was inferred by constructing a neighbor-joining tree of concatenated nucleotide sequences of the five loci (ITS, MS1, MS3, MS4, and MS7) using the program Mega 5[5] based on the evolutionary distances calculated by Kimura 2-parameter model. The reliability of cluster formation was assessed by the bootstrap analysis with 1,000 replicates. Median-joining analysis implemented in the software network version 4.6.1.0[6] was conducted with the default parameters to identify a substructure of the *E. bieneusi* isolates.

RESULTS

Genetic Polymorphism and MLGs

At the four micro-/mini-satellite loci (MS1, MS3, MS4, and MS7), 153 (50.16%), 131 (42.95%), 133 (43.61%), and 128 (41.97%) DNA specimens were amplified and sequenced successfully, respectively. 44, 17, 26, and 24 genotypes were identified at

[1]http://www.ncbi.nlm.nih.gov/blast/

[2]http://www.clustal.org/

[3]http://guanine.evolbio.mpg.de/cgi-bin/lian/lian.cgi.pl

[4]http://www.ub.es/dnasp/

[5]http://www.megasoftware.net/

[6]http://www.fluxus-engineering.com/sharenet_rn.htm

TABLE 2 | Multilocus linkage disequilibrium analysis based on allelic data of *E. bieneusi*.

Population group	No.	H	I^S_A	P_{MC}	V_D	L	$V_D > L$	LD or LE
ITS included	90	0.8032 ± 0.0366	0.5735	$<1.00 \times 10^{-02}$	2.5154	0.8363	Yes	LD
ITS excluded	90	0.8203 ± 0.0417	0.5129	$<1.00 \times 10^{-02}$	1.4436	0.6009	Yes	LD
Subgroup 1	52	0.5306 ± 0.1008	0.4432	$<1.00 \times 10^{-02}$	2.8890	1.2536	Yes	LD
Subgroup 2	31	0.6935 ± 0.0518	0.4705	$<1.00 \times 10^{-02}$	2.9080	1.1424	Yes	LD
Subgroup 3	7	0.7714 ± 0.0614	0.3177	$<1.00 \times 10^{-02}$	1.8307	1.2381	Yes	LD
ITS included[a]	48	0.8898 ± 0.0218	0.3647	$<1.00 \times 10^{-02}$	1.1821	0.5234	Yes	LD
ITS excluded[a]	46	0.9026 ± 0.0240	0.2695	$<1.00 \times 10^{-02}$	0.6232	0.3732	Yes	LD
Subgroup 1[a]	23	0.7043 ± 0.0780	0.2788	$<1.00 \times 10^{-02}$	1.9450	1.1813	Yes	LD
Subgroup 2[a]	18	0.7614 ± 0.0693	0.2435	$<1.00 \times 10^{-02}$	1.6037	1.0272	Yes	LD

No identical MLGs were found in subgroup 3. [a]Considering each group of specimens with the same MLG as one individual.

TABLE 3 | Genetic diversity in *E. bieneusi* isolates (n = 90) based on analysis of concatenated multilocus sequences.

Concatenated sequence	No. of site (bp)	No. of Polymorphic (segregating) sites	No. of haplotypes, H	Gene diversity, Hd	Interlocus genetic association (*Zns*)	95% CI (*Zns*)	Intragenic linkage disequilibrium (LD) |D′|	Recombination events (Rms)
ITS included	2584	118	41	0.942	0.2217	0.05117–0.37500	Y = 0.9750 – 0.0085X	14
ITS excluded	2341	90	39	0.940	0.1766	0.04976–0.34890	Y = 0.9857 – 0.0279X	12

|D′|: linkage disequilibrium (LD), where Y is the LD value, and X is the nucleotide distance in kb.

the MS1, MS3, MS4, and MS7 loci, respectively. SNPs and repeats could be observed: trinucleotide repeats of TGC (n = 1), TAA (n = 5–38) and T(C)AC(T) (n = 3–10) at the MS1 locus, dinucleotide repeats of TA (n = 4–17) at the MS3 locus, tetranucleotide repeats of G(A)G(A)TA (n = 1–2) at the MS4 locus, and trinucleotide repeats of TAA (n = 3–11) at the MS7 locus. Additionally, INDELs were observed at the MS4 locus (**Table 1**). The nucleotide sequences of all the genotypes were deposited in the GenBank under following accession numbers: MT267413–MT267456 (MS1), MT267457–MT267473 (MS3), MT267363–MT267388 (MS4) and MT267389–MT267412 (MS7).

In the present study, a total of 90 ITS-positive DNA specimens were amplified and sequenced successfully at the MS1, MS3, MS4, and MS7 loci, producing 48 MLGs based on concatenated sequences of all the five genetic loci including ITS (**Supplementary Table S4**).

LD and Population Structure

All the 90 DNA specimens genotyped successfully at all the five genetic loci including ITS were used in LD analysis. Multilocus LD was assessed by the calculation of I^S_A based on the allelic profile data of the five genetic loci including ITS using software LIAN 3.5. The value of I^S_A (0.5735) was positive and the V_D (2.5154) was more than L (0.8363). A Monte Carlo analysis led to the generation of a significant P_{MC} value (<0.01). These above results showed the population of *E. bieneusi* in the investigated areas was in LD with clonality. To exclude the possibility that LD could be attributable to a clonal expansion of one or more MLGs, which could mask the underlying equilibrium, the same MLG was scored as one individual and multilocus LD was assessed again. LD was still existed based on the fact that the I^S_A value was still above zero (I^S_A = 0.3647), and V_D (1.1821) was more than L (0.5234), suggesting a clonal population structure. As ITS locus has a higher degree of genetic diversity than other genetic loci in *E. bieneusi* genome, the analysis was conducted on allelic profile data excluding ITS. Similar results were obtained, indicating ITS sequences had little influence on the result of LD analysis (**Table 2**).

The software DnaSP5.10.01 was used to investigate the genetic diversity of *E. bieneusi* isolates. 118 and 90 polymorphic/segregating sites were observed in the concatenated sequence data of five genetic loci including ITS (2584 bp in length) and four genetic loci excluding ITS (2341 bp in length), respectively. LD was also estimated using the Z_nS statistics. However, LD was incomplete based on the |D′| Y = 0.9750 – 0.0085X (including ITS) and the |D′| Y = 0.9857 – 0.0279X (excluding ITS). A negative slope implied a decline in LD with increasing nucleotide distance and the potential occurrence of recombination. Indeed, limited recombination events were detected in analysis of concatenated multilocus sequences (**Table 3**).

Population Substructure

Phylogenetic relationship of the *E. bieneusi* isolates was inferred by neighbor-joining analysis of the concatenated sequences of all the five genetic loci. Two phylogenetic clusters were

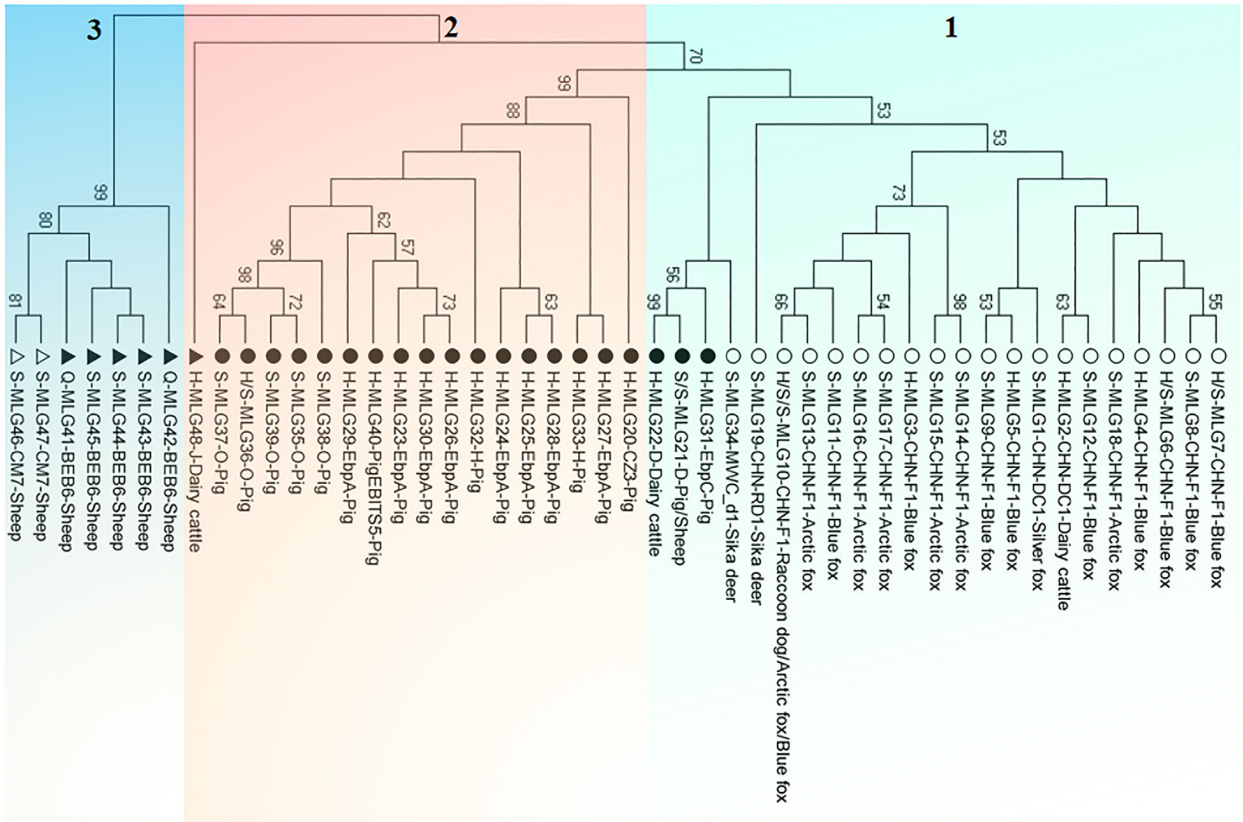


FIGURE 2 | Phylogenetic relationship of multilocus genotypes (MLGs) of animal-derived *E. bieneusi* isolates. The relationship was inferred by a neighbor-joining analysis of 48 concatenated nucleotide sequences of the five genetic loci (ITS, MS1, MS3, MS4, MS7) from 90 *E. bieneusi* isolates belonging to 14 ITS genotypes (CHN-DC1, CHN-F1, CHN-RD1, CZ3, D, EbpA, CHN-RD1, CZ3, D, EbpA, PigEBITS5, BEB6, CM7, and J) from dairy cattle, pigs, sheep, sika deer, blue foxes, arctic foxes, a silver fox, and a raccoon dog (**Supplementary Table S4**) based on genetic distance by the Kimura 2-parameter model. Each sequence is identified by its sampling site (H: Harbin; Q: Qiqihar; S: Suihua), MLG designation and ITS genotype as well as host origin. The solid and open circles/triangles indicate zoonotic and host-adapted ITS genotypes, respectively, with circle and triangle ones belonging to ITS group 1 and ITS group 2, respectively.

observed, and the larger one was further divided into two subdivisions, leading to three subgroups altogether, reliability and confidence of which were verified based on the calculation of pairwise genetic distance (F_{ST}) and gene flow (Nm) using DnaSP (**Supplementary Table S5**). Subgroup 1 was composed of MLGs from ITS genotypes CHN-DC1, CHN-F1, CHN-RD1, D, EbpC and MWC_d1 (in ITS group 1), mainly isolated from fur animals (blue foxes, arctic foxes, silver foxes and raccoon dogs). Subgroup 2 comprised MLGs from ITS genotypes CZ3, EbpA, H, O and PigEBITS5 (in ITS group 1) and J (in ITS group 2), which were mainly from pigs. Subgroup 3 consisted of MLGs from ITS genotype BEB6 and CM7 (in ITS group 2) derived from sheep. There was an absence of geographical segregation in *E. bieneusi* population in the present study (**Figure 2**). Meanwhile, the MLGs from zoonotic ITS genotypes were generally separated from the MLGs from host-adapted ones in ITS group 1 (**Figure 2**), which was also supported by median-joining network analysis of the entire concatenated multilocus sequences (**Figure 3**).

Subgroups 1–3 contained 23, 18, and seven MLGs derived from 52, 31 and seven *E. bieneusi* isolates, respectively. The I^S_A values were positive for each of the three subgroups by LD analysis of the allelic profile data. Meanwhile, V_D values larger than the L values for subgroups 1–3 suggested the existence of LD and clonal population structures, with significant LD ($P_{MC} < 0.01$) for subgroups 1–3. Multilocus LD was recalculated when the same MLG treated as one individual. Subgroups 1, 2 both had positive I^S_A values and $V_D > L$ with significant P_{MC}, confirming the existence of LD and clonal population structures. Due to no identical MLGs found in subgroup 3, the subgroup was not assessed again (**Table 2**).

DISCUSSION

MLST analysis has been widely used to study population genetics of *E. bieneusi* (Li and Xiao, 2019). In the present study, MLST analysis of the MS1, MS3, MS4, and MS7 loci was performed

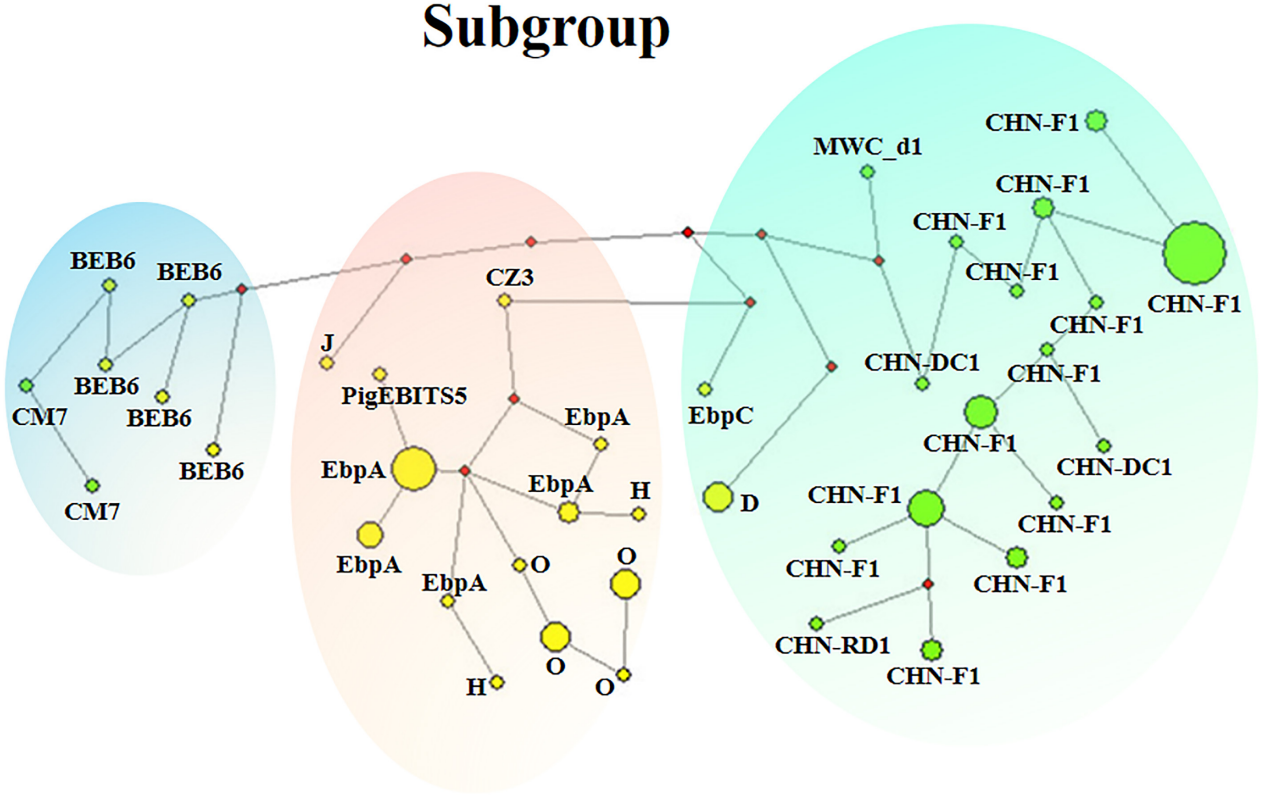


FIGURE 3 | Median-joining network analysis for inferring intraspecific phylogenies of 90 *E. bieneusi* isolates belonging to 14 ITS genotypes from eight mammal species. The yellow and green colors in circles represent zoonotic ITS genotypes and host-adapted ITS genotypes, respectively. The diameter of each circle is proportional to the frequency of each multilocus genotype (MLG) on the basis of polymorphic/segregating sites of concatenated sequences of the five genetic loci including ITS. Length of lines connecting MLGs is proportional to the number of single nucleotide polymorphisms. ITS genotypes are labeled besides the circles. Three clusters marked in green, pink and blue are corresponded to subgroups 1–3, respectively.

on 305 *E. bieneusi* isolates derived from various animal species. Differences in PCR amplification efficiencies were observed at the four different MLST loci: 50.16% at the MS1, 42.95% at the MS3, 43.61% at the MS4 and 41.97% at the MS7, which were all higher than those of *E. bieneusi* isolates in China from horses (6.67–18.67%) (Deng et al., 2016) and from Asiatic black bears (27.59–31.03%) (Deng et al., 2017), but lower than those of *E. bieneusi* isolates from NHPs (77.78–86.51%) (Karim et al., 2014) and from captive wild animals (76.74–86.05%) (Li et al., 2016a). Similar amplification efficiencies were found in one study of multilocus genotyping of *E. bieneusi* isolates from calves in China (38.36–61.64%) (Wang et al., 2016). The inconsistency of amplification efficiencies might be related to different ITS genotypes of *E. bieneusi* isolates in MLST analysis. In fact, micro-/mini-satellite sequences were searched in the whole-genome sequence database of the human-derived *E. bieneusi* isolate (H348), and the present PCR primers were designed for target genes (MS1, MS3, MS4, and MS7) based on nucleotide sequences flanking the potential micro-/mini-satellite repeats (Feng et al., 2011). Li et al. (2019b) has ever pointed out that hypermutation in *E. bieneusi* genome possibly hinders some isolates from effectively amplifying. Therefore, it is necessary to develop additional reliable and effective genetic markers that will successfully amplify isolates of all *E. bieneusi* genotypes to better understand their host specificity and population structure.

In the present study, 44 genotypes were obtained at the MS1 locus, followed by 26, 24, and 17 genotypes at the MS4, MS7, and MS3, respectively. Previous studies described different numerical ordinations of genotypes at the four loci (Feng et al., 2011; Li et al., 2013, 2017; Karim et al., 2014; Deng et al., 2016; Wan et al., 2016; Wang et al., 2018a, 2019; Wu et al., 2018). However, MS1 was generally observed to have the highest resolution in most studies, leading to more genotypes at this locus than the MS3, MS4, and MS7 loci. The appearance of this phenomenon might be related to the number and host origin of *E. bieneusi* isolates analyzed and the constitution of ITS genotypes.

In the present study, a total of 48 MLGs were identified out of 90 ITS-positive *E. bieneusi* isolates. The presence of significant LD and a low number of recombination events indicated a clonal population structure of *E. bieneusi* in the investigated areas. The results imply that MLGs of *E. bieneusi* are relatively stable in time and in space in the investigated areas, and MLST analysis can be used effectively in the longitudinal tracking of the transmission of *E. bieneusi* in the community and in investigation of outbreaks. To date, at least 11 studies have assessed the population structure of *E. bieneusi*, in which over 800 *E. bieneusi* isolates have been

used for LD analysis, including over 100 human-derived isolates belonging to 14 ITS genotypes from Peru, India and Nigeria and over 700 animal-derived isolates belonging to 60 ITS genotypes from China and Kenya, and LD analysis revealed the wide existence of a clonal population structure (Li et al., 2012, 2013, 2016b, 2017, 2019; Karim et al., 2014; Wan et al., 2016; Wang et al., 2018a, 2019; Zhang et al., 2020; this study) (**Table 4**).

Based on the present result of phylogenetic analysis of MLGs, genetic segregation related to hosts was observed at a large extent in the *E. bieneusi* population from various mammal species, with subgroups 1–3 composed of MLGs mainly from fur animals, pigs and sheep, respectively. In fact, host adaptation has been reported in some studies of *E. bieneusi* populations: fur animals (foxes and raccoon dogs) versus humans (Li et al., 2016b); pigs versus others (humans, NHPs, cattle, horses, and bears) (Li et al., 2019); dairy cattle versus others (humans, NHPs, pigs, bears, horses and kangaroos) (Wang et al., 2019). These results will be helpful to understand zoonotic or interspecies transmission of *E. bieneusi*.

In the present study, the *E. bieneusi* MLGs were not observed to be related to geographical locations of DNA specimens. This result was similar to some previous studies of *E. bieneusi* populations from humans in Peru, India and Nigeria (Li et al.,

TABLE 4 | *Enterocytozoon bieneusi* isolates used for LD analysis and characterizations of population structure and substructure.

Host (country)	*E. bieneusi* isolates used for LD analysis		No. of MLGs identified	Characterization		References
	Number	ITS genotypes		Population	Subgroup (n)	
Human (Peru)	72	A, D, EbpC, Peru7, Peru8, Peru10, Peru11, Type IV, WL11	39	Clonal	Clonal (n = 1); epidemic (n = 1)	Li et al., 2012
Human (Peru, India, Nigeria), Olive baboon (Kenya)	64[b]	A, D, EbpC, IH, Nig2, Nig3, Nig5, Peru7, Peru8, Peru10, Peru11, PigEBITS7, Type IV, WL11	48	Clonal[d]	Clonal (n = 1); epidemic (n = 1)	Li et al., 2013
NHP (China)	85	CM2, D, Henan-V, Macaque3, Peru8, Peru11, PigEBITS7, Type IV	59	Clonal	Epidemic (n = 4)	Karim et al., 2014
Fox, raccoon dog[a] (China)	39	D	10	Clonal	NA	Li et al., 2016b
Pig (China)	101	CHN7, CS-4, EbpA, EbpB, EbpC, LW1, Henan-IV, O, PigEBITS3	44	Clonal	Clonal (n = 1); epidemic (n = 2)	Wan et al., 2016
	109[c]	CHC5, CHG3, CHN7, CS-4, D, EbpA, H, Henan-IV, PigEb4, PigEBITS4, PigEBITS5, SHZA1, SHZC1, SLTC1-SLTC3, SMXB1, SMXC1, SMXD1, SMXD2, SYLA1-SYLA5, SYLC1, SYLD1, SZZA1, SZZA2, SZZB1, SZZC1, SZZD1, SZZD2	87	Clonal	NA	Wang et al., 2018a
	93	EbpC, EbpA, H, PigEBITS4	76	Clonal	NA	Li et al., 2019
	52	CM11, D, EbpC, EbpA, FJF	48	Clonal	NA	Zhang et al., 2020
Dairy cattle (China)	106	BEB4, I, J	71	Clonal	NA	Wang et al., 2019
Giant panda (China)	36	D, SC02, SC05, SC06	24	Clonal	Clonal (n = 1); epidemic (n = 2)	Li et al., 2017
Dairy cattle, pig, sheep, sika deer, blue fox, arctic fox, silver fox, raccoon dog (China)	90	BEB6, CHN-DC1, CHN-F1, CHN-RD1, CM7, CZ3, D, EbpA, EbpC, H, J, MWC_d1, O, PigEBITS5	48	Clonal	Clonal (n = 3); epidemic (n = 1)	This study

*Five genotypes in **Table 4** have been changed to the first published names instead of genotype names described in original papers: CM1 to Macaque3 (Karim et al., 2014); Henan-I to LW1 (Wan et al., 2016); CHG19 to PigEBITS4 (Wang et al., 2018a; Li et al., 2019); CHS5 to EbpA (Wang et al., 2018a; Zhang et al., 2020); CM6 to CS-4 (Wang et al., 2018a).* [a]*39 E. bieneusi isolates are derived from foxes (n = 30) and raccoon dogs (n = 9).* [b]*64 E. bieneusi isolates used for LD analysis are derived from humans (n = 59) in Peru, India, Nigeria and olive baboons (n = 5) in Kenya. Among them, MLST data of 26 human-derived isolates are from a study conducted in Peru by Li et al. (2012).* [c]*These ITS genotypes are used for MLST analysis without specific information of genotypes used for LD analysis.* [d]*A clonal population structure is observed in each of 3 geographical locations (Peru, India, and Nigeria).*

2013) and from animals in China (Li et al., 2017; Wang et al., 2019; Zhang et al., 2020). However, genetic segregation at some geographical level appeared in the other studies conducted in China (Wan et al., 2016; Wang et al., 2018a; Li et al., 2019). No geographical segregation of *E. bieneusi* isolates was seen in the investigated areas, which might be related to smaller geographical areas and number of *E. bieneusi* isolates as well as fewer ITS genotypes analyzed.

To date, approximately 500 *E. bieneusi* ITS genotypes have been identified: ∼60 only found in humans; ∼390 only found in animals; ∼50 found in both humans and animals (Li et al., 2019b; Shen et al., 2020). However, the genetic reasons behind this observation of *E. bieneusi* genotypes remain unclear, especially for those genotypes belonging to group 1 with zoonotic nature and group 2 with increasing importance for public health. In the present study, 48 MLGs belonging to 14 ITS genotypes were divided into three different subgroups. The MLGs were generally separated by zoonotic potential, revealing possible genetic segregation of *E. bieneusi* population. Subgroup 1 was composed of the MLGs from zoonotic ITS genotypes D and EbpC with wide host ranges and those from host-adapted ITS genotypes (CHN-DC1, CHN-F1, CHN-RD1 and MWC_d1) with narrow host ranges based on the summarized data in **Supplementary Table S6**. Meanwhile, the MLG from ITS genotype J fell into subgroup 2 consisting of those from zoonotic ITS genotypes CZ3, EbpA, H, O and PigEBITS5. Combined with network analysis, subgroups 1, 3 might originate from subgroup 2, with D and EbpC, and J as evolutionary transitions, respectively. It was speculated that there might be different *E. bieneusi* populations with different genetic structures and public health potential in their evoluting processes, with the possible direction from zoonotic genotypes to host-adapted genotypes. Finally, some genotypes of *E. bieneusi* might be adapted to changing host environments, and were successfully chosen by hosts. Nevertheless, it is still unclear on the possible factors promoting adaptation of different *E. bieneusi* genotypes to specific hosts or host changes.

CONCLUSION

The present study describes the genetic characterizations of *E. bieneusi* isolates derived from various mammal and bird species at the four MLST loci (MS1, MS3, MS4, and MS7) in northeastern China. 48 MLGs were produced out of 14 ITS genotypes, revealing MLGs have a higher genetic diversity than ITS genotypes. MLST data and a clonal population structure of *E. bieneusi* will aid in the longitudinal tracking of the attribution of source of infection/contamination and in elucidating transmission dynamics, and will provide valuable information for making efficient control strategies to intervene with and prevent occurrence of microsporidiosis caused by *E. bieneusi* among animals and transmission of *E. bieneusi* from animals to humans in the investigated areas. The result of host adaptation of *E. bieneusi* population has a reference effect on further understanding of zoonotic or interspecies transmission of *E. bieneusi*. The observation of genetic segregation of *E. bieneusi* population by zoonotic potential and possible evolution direction of *E. bieneusi* from zoonotic genotypes to host-adapted genotypes will improve our understanding of evolution of *E. bieneusi* population. However, the role of host, geographical, and temporal factors in the evolution mechanism of *E. bieneusi* populations needs to be characterized in the future.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because all the specimens analyzed were DNA preparations from animal feces.

AUTHOR CONTRIBUTIONS

AL and YS designed this study and made the final revision. XL and YW performed the experiments, wrote the first draft of the manuscript, and prepared the tables and figures. XL, YW, BG, FY, YJ, and KZ analyzed the data. WZ and JC contributed to the reagents and materials. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported partially by the National Science and Technology Major Program of China (No. 2018ZX10713001-004 to YS), the National Key Research and Development Program of China (No. 2017YFD0501300 to YJ), and the Natural Science Foundation of Heilongjiang Province of China (No. H2017006 to FY).

ACKNOWLEDGMENTS

We thank the staff at the Department of Parasitology, Harbin Medical University for their assistance with ITS genotype identification of DNA specimens of *E. bieneusi*.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01406/full#supplementary-material

REFERENCES

Buckholt, M. A., Lee, J. H., and Tzipori, S. (2002). Prevalence of *Enterocytozoon bieneusi* in swine: an 18-month survey at a slaughterhouse in Massachusetts. *Appl. Environ. Microbiol.* 68, 2595–2599. doi: 10.1128/AEM.68.5.2595-2599

Deng, L., Li, W., Zhong, Z., Gong, C., Cao, X., Song, Y., et al. (2017). Multilocus genotypes of *Enterocytozoon bieneusi* in captive Asiatic black bears in southwestern China: high genetic diversity, broad host range, and zoonotic potential. *PLoS One* 12:e0171772. doi: 10.1371/journal.pone.0171772

Deng, L., Li, W., Zhong, Z., Gong, C., Liu, X., Huang, X., et al. (2016). Molecular characterization and multilocus genotypes of *Enterocytozoon bieneusi* among horses in southwestern China. *Parasit. Vectors* 9:561. doi: 10.1186/s13071-016-1844-3

Feng, Y., Li, N., Dearen, T., Lobo, M. L., Matos, O., Cama, V., et al. (2011). Development of a multilocus sequence typing tool for high-resolution genotyping of *Enterocytozoon bieneusi*. *Appl. Environ. Microbiol.* 77, 4822–4828. doi: 10.1128/AEM.02803-10

Gong, B., Yang, Y., Liu, X., Cao, J., Xu, M., Xu, N., et al. (2019). First survey of *Enterocytozoon bieneusi* and dominant genotype Peru6 among ethnic minority groups in southwestern China's Yunnan Province and assessment of risk factors. *PLoS Negl. Trop. Dis.* 13:e0007356. doi: 10.1371/journal.pntd.0007356

Haubold, B., Travisano, M., Rainey, P. B., and Hudson, R. R. (1998). Detecting linkage disequilibrium in bacterial populations. *Genetics* 150, 1341–1348.

Karim, M. R., Wang, R., He, X., Zhang, L., Li, J., Rume, F. I., et al. (2014). Multilocus sequence typing of *Enterocytozoon bieneusi* in nonhuman primates in China. *Vet. Parasitol.* 200, 13–23. doi: 10.1016/j.vetpar.2013.12.004

Li, D., Zheng, S., Zhou, C., Karim, M. R., Wang, L., Wang, H., et al. (2019). Multilocus Typing of *Enterocytozoon bieneusi* in Pig Reveals the High Prevalence, Zoonotic Potential, Host Adaptation and Geographical Segregation in China. *J. Eukaryot. Microbiol.* 66, 707–718. doi: 10.1111/jeu.12715

Li, W., Cama, V., Akinbo, F. O., Ganguly, S., Kiulia, N. M., Zhang, X., et al. (2013). Multilocus sequence typing of *Enterocytozoon bieneusi*: lack of geographic segregation and existence of genetically isolated sub-populations. *Infect. Genet. Evol.* 14, 111–119. doi: 10.1016/j.meegid.2012.11.021

Li, W., Cama, V., Feng, Y., Gilman, R. H., Bern, C., Zhang, X., et al. (2012). Population genetic analysis of *Enterocytozoon bieneusi* in humans. *Int. J. Parasitol.* 42, 287–293. doi: 10.1016/j.ijpara.2012.01.003

Li, W., Feng, Y., and Santin, M. (2019a). Host Specificity of *Enterocytozoon bieneusi* and Public Health Implications. *Trends Parasitol.* 35, 436–451. doi: 10.1016/j.pt.2019.04.004

Li, W., Feng, Y., Zhang, L., and Xiao, L. (2019b). Potential impacts of host specificity on zoonotic or interspecies transmission of *Enterocytozoon bieneusi*. *Infect. Genet. Evol.* 75:104033. doi: 10.1016/j.meegid.2019.104033

Li, W., Song, Y., Zhong, Z., Huang, X., Wang, C., Li, C., et al. (2017). Population genetics of *Enterocytozoon bieneusi* in captive giant pandas of China. *Parasit. Vectors* 10:499. doi: 10.1186/s13071-017-2459-z

Li, W., Deng, L., Yu, X., Zhong, Z., Wang, Q., Liu, X., et al. (2016a). Multilocus genotypes and broad host-range of *Enterocytozoon bieneusi* in captive wildlife at zoological gardens in China. *Parasit. Vectors* 9:395. doi: 10.1186/s13071-016-1668-1

Li, W., Wan, Q., Yu, Q., Yang, Y., Tao, W., Jiang, Y., et al. (2016b). Genetic variation of mini- and microsatellites and a clonal structure in *Enterocytozoon bieneusi* population in foxes and raccoon dogs and population differentiation of the parasite between fur animals and humans. *Parasitol. Res.* 115, 2899–2904. doi: 10.1007/s00436-016-5069-3

Li, W., and Xiao, L. (2019). Multilocus sequence typing and population genetic analysis of *Enterocytozoon bieneusi*: host specificity and its impacts on public health. *Front. Genet.* 10:307. doi: 10.3389/fgene.2019.00307

Liu, W., Nie, C., Zhang, L., Wang, R., Liu, A., Zhao, W., et al. (2015). First detection and genotyping of *Enterocytozoon bieneusi* in reindeers (*Rangifer tarandus*): a zoonotic potential of ITS genotypes. *Parasit. Vectors* 8:526. doi: 10.1186/s13071-015-1155-0

Matos, O., Lobo, M. L., and Xiao, L. (2012). Epidemiology of *Enterocytozoon bieneusi* Infection in Humans. *J. Parasitol. Res.* 2012:981424. doi: 10.1155/2012/981424

Qiu, L., Xia, W., Li, W., Ping, J., Ding, S., and Liu, H. (2019). Author Correction: the prevalence of microsporidia in China: a systematic review and meta-analysis. *Sci. Rep.* 9:18482. doi: 10.1038/s41598-019-54376-8

Santín, M., and Fayer, R. (2009). *Enterocytozoon bieneusi* genotype nomenclature based on the internal transcribed spacer sequence: a consensus. *J. Eukaryot. Microbiol.* 56, 34–38. doi: 10.1111/j.1550-7408.2008.00380.x

Santín, M., and Fayer, R. (2011). Microsporidiosis: *Enterocytozoon bieneusi* in domesticated and wild animals. *Res. Vet. Sci.* 90, 363–371. doi: 10.1016/j.rvsc.2010.07.014

Shen, Y., Gong, B., Liu, X., Wu, Y., Yang, F., Xu, J., et al. (2020). First identification and genotyping of *Enterocytozoon bieneusi* in humans in Myanmar. *BMC Microbiol.* 20:10. doi: 10.1186/s12866-019-1694-1

Wan, Q., Xiao, L., Zhang, X., Li, Y., Lu, Y., Song, M., et al. (2016). Clonal Evolution of *Enterocytozoon bieneusi* populations in swine and genetic differentiation in subpopulations between isolates from swine and humans. *PLoS Negl. Trop. Dis.* 10:e0004966. doi: 10.1371/journal.pntd.0004966

Wang, H. Y., Qi, M., Sun, M. F., Li, D. F., Wang, R. J., Zhang, S. M., et al. (2019). Prevalence and Population Genetics Analysis of *Enterocytozoon bieneusi* in dairy cattle in China. *Front. Microbiol.* 10:1399. doi: 10.3389/fmicb.2019.01399

Wang, S. S., Li, J. Q., Li, Y. H., Wang, X. W., Fan, X. C., Liu, X., et al. (2018a). Novel genotypes and multilocus genotypes of *Enterocytozoon bieneusi* in pigs in northwestern China: a public health concern. *Infect. Genet. Evol.* 63, 89–94. doi: 10.1016/j.meegid.2018.05.015

Wang, S. S., Wang, R. J., Fan, X. C., Liu, T. L., Zhang, L. X., and Zhao, G. H. (2018b). Prevalence and genotypes of *Enterocytozoon bieneusi* in China. *Acta Trop.* 183, 142–152. doi: 10.1016/j.actatropica.2018.04.017

Wang, X. T., Wang, R. J., Ren, G. J., Yu, Z. Q., Zhang, L. X., Zhang, S. Y., et al. (2016). Multilocus genotyping of *Giardia duodenalis* and *Enterocytozoon bieneusi* in dairy and native beef (Qinchuan) calves in Shaanxi province, northwestern China. *Parasitol. Res.* 15, 1355–1361. doi: 10.1007/s00436-016-4908-6

Widmer, G., and Akiyoshi, D. E. (2010). Host-specific segregation of ribosomal nucleotide sequence diversity in the microsporidian *Enterocytozoon bieneusi*. *Infect. Genet. Evol.* 10, 122–128. doi: 10.1016/j.meegid.2009.11.009

Wu, J., Han, J. Q., Shi, L. Q., Zou, Y., Li, Z., Yang, J. F., et al. (2018). Prevalence, genotypes, and risk factors of *Enterocytozoon bieneusi* in Asiatic black bear (*Ursus thibetanus*) in Yunnan Province, Southwestern China. *Parasitol. Res.* 117, 1139–1145. doi: 10.1007/s00436-018-5791-0

Yang, Z., Zhao, W., Shen, Y., Zhang, W., Shi, Y., Ren, G., et al. (2016). Subtyping of *Cryptosporidium cuniculus* and genotyping of *Enterocytozoon bieneusi* in rabbits in two farms in Heilongjiang Province. *China. Parasite* 23:52. doi: 10.1051/parasite/2016063

Zhang, N., Wu, R., Ji, T., Cui, L. L., Cao, H. X., Li, D., et al. (2020). Molecular Detection, Multilocus Genotyping, and Population Genetics of *Enterocytozoon bieneusi* in Pigs in Southeastern China. *J. Eukaryot. Microbiol.* 67, 107–114. doi: 10.1111/jeu.12759

Zhang, X., Wang, Z., Su, Y., Liang, X., Sun, X., Peng, S., et al. (2011). Identification and genotyping of *Enterocytozoon bieneusi* in China. *J. Clin. Microbiol.* 49, 2006–2008. doi: 10.1128/JCM.00372-11

Zhang, X. X., Cong, W., Lou, Z. L., Ma, J. G., Zheng, W. B., Yao, Q. X., et al. (2016). Prevalence, risk factors and multilocus genotyping of *Enterocytozoon bieneusi* in farmed foxes (*Vulpes lagopus*). Northern China. *Parasit. Vectors* 9:72. doi: 10.1186/s13071-016-1356-1

Zhao, W., Wang, J., Ren, G., Yang, Z., Yang, F., Zhang, W., et al. (2018). Molecular characterizations of *Cryptosporidium* spp. and *Enterocytozoon bieneusi* in brown rats (*Rattus norvegicus*) from Heilongjiang Province. China. *Parasit. Vectors* 11:313. doi: 10.1186/s13071-018-2892-7

Zhao, W., Wang, J., Yang, Z., and Liu, A. (2017). Dominance of the *Enterocytozoon bieneusi* genotype BEB6 in red deer (*Cervus elaphus*) and Siberian roe deer (*Capreolus pygargus*) in China and a brief literature review. *Parasite* 24:54. doi: 10.1051/parasite/2017056

Zhao, W., Yu, S., Yang, Z., Zhang, Y., Zhang, L., Wang, R., et al. (2016). Genotyping of *Enterocytozoon bieneusi* (Microsporidia) isolated from various birds in China. *Infect. Genet. Evol.* 40, 151–154. doi: 10.1016/j.meegid.2016.02.037

Zhao, W., Zhang, W., Yang, F., Zhang, L., Wang, R., Cao, J., et al. (2015a). *Enterocytozoon bieneusi* in Dairy Cattle in the Northeast of China: genetic diversity of ITS gene and evaluation of zoonotic transmission potential. *J. Eukaryot. Microbiol.* 62, 553–560. doi: 10.1111/jeu.12210

Zhao, W., Zhang, W., Yang, D., Zhang, L., Wang, R., and Liu, A. (2015b). Prevalence of *Enterocytozoon bieneusi* and genetic diversity of ITS genotypes in sheep and goats in China. *Infect. Genet. Evol.* 32, 265–270. doi: 10.1016/j.meegid.2015a.03.026

Zhao, W., Zhang, W., Yang, Z., Liu, A., Zhang, L., Yang, F., et al. (2015c). Correction: genotyping of *Enterocytozoon bieneusi* in Farmed Blue Foxes (*Alopex lagopus*) and Raccoon Dogs (*Nyctereutes procyonoides*) in China. *PLoS One* 10:e0143992. doi: 10.1371/journal.pone.0143992

Zhao, W., Zhang, W., Wang, R., Liu, W., Liu, A., Yang, D., et al. (2014a). *Enterocytozoon bieneusi* in sika deer (*Cervus nippon*) and red deer (*Cervus elaphus*): deer specificity and zoonotic potential of ITS genotypes. *Parasitol. Res.* 113, 4243–4250. doi: 10.1007/s00436-014-4100-9

Zhao, W., Zhang, W., Yang, F., Cao, J., Liu, H., Yang, D., et al. (2014b). High prevalence of *Enterocytozoon bieneusi* in asymptomatic pigs and assessment of zoonotic risk at the genotype level. *Appl. Environ. Microbiol.* 80, 3699–3707. doi: 10.1128/AEM.00807-14

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.

Copyright © 2020 Liu, Wu, Yang, Gong, Jiang, Zhou, Cao, Zhang, Liu and Shen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 August 2020
doi: 10.3389/fmicb.2020.01930

New *Plasmodium vivax* Genomes From the China-Myanmar Border

*Awtum M. Brashear[1], Adam C. Huckaby[2], Qi Fan[3], Luke J. Dillard[2], Yubing Hu[4], Yuling Li[4], Yan Zhao[4], Zenglei Wang[5], Yaming Cao[4], Jun Miao[1], Jennifer L. Guler[2] and Liwang Cui[1]**

[1] Department of Internal Medicine, University of South Florida, Tampa, Tampa, FL, United States, [2] Department of Biology, University of Virginia, Charlottesville, VA, United States, [3] Dalian Institute of Science and Technology, Dalian, China, [4] Department of Immunology, College of Basic Medical Science, China Medical University, Shenyang, China, [5] NHC Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Plasmodium vivax is increasingly the dominant species of malaria in the Greater Mekong Subregion (GMS), which is pursuing regional malaria elimination. *P. vivax* lineages in the GMS are poorly characterized. Currently, *P. vivax* reference genomes are scarce due to difficulties in culturing the parasite and lack of high-quality samples. In addition, *P. vivax* is incredibly diverse, necessitating the procurement of reference genomes from different geographical regions. Here we present four new *P. vivax* draft genomes assembled *de novo* from clinical samples collected in the China-Myanmar border area. We demonstrate comparable length and content to existing genomes, with the majority of structural variation occurring around subtelomeric regions and exported proteins, which we corroborated with detection of copy number variations in these regions. We predicted peptides from all *PIR* gene subfamilies, except for *PIR* D. We confirmed that proteins classically labeled as *PIR* D family members are not identifiable by PIR motifs, and actually bear stronger resemblance to DUF (domain of unknown function) family DUF3671, potentially pointing to a new, closely related gene family. Further, phylogenetic analyses of *MSP7* genes showed high variability within the *MSP7-B* family compared to *MSP7-A* and *-C* families, and the result was comparable to that from whole genome analyses. The new genome assemblies serve as a resource for studying *P. vivax* within the GMS.

Keywords: *Plasmodium vivax*, malaria, China-Myanmar border, genome assembly, next-generation sequencing

OPEN ACCESS

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

Reviewed by:
Karine Gaelle Le Roch,
University of California, Riverside, United States
Jianbing Mu,
National Institutes of Health (NIH), United States

****Correspondence:***
Liwang Cui
liwangcui@usf.edu

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

Received: *31 May 2020*
Accepted: *22 July 2020*
Published: *11 August 2020*

Citation:
Brashear AM, Huckaby AC, Fan Q, Dillard LJ, Hu Y, Li Y, Zhao Y, Wang Z, Cao Y, Miao J, Guler JL and Cui L (2020) New Plasmodium vivax Genomes From the China-Myanmar Border. Front. Microbiol. 11:1930. doi: 10.3389/fmicb.2020.01930

INTRODUCTION

The Greater Mekong Subregion (GMS) is problematic in the grand scheme of malaria control due to the repeated emergence of antimalarial resistance in this region (Cui et al., 2012; WHO, 2016; Imwong et al., 2017). The World Health Organization has declared a goal to eliminate malaria in the GMS by 2030 (WHO, 2016). Due to the uneven distribution of malaria along international borders within the GMS, cross-border transmission of malaria from areas with high endemicity to malaria-free areas is a concern (Lo et al., 2015, 2017; Kittichai et al., 2017). Myanmar is one of the six countries within the GMS with the highest malaria burden (Cui et al., 2012; WHO, 2018). The China-Myanmar border is an example of an area wherein cross-border migration is particularly

concerning and can result in outbreaks in the less endemic area (Duan et al., 2013; Lo et al., 2015, 2017).

As of 2016, *Plasmodium vivax* was responsible for between 40 and 60% of malaria cases in the GMS countries, and the vast majority of malaria on the China-Myanmar border (WHO, 2018; Geng et al., 2019). *Plasmodium vivax* has a hypnozoite stage which will go dormant for long periods of time before re-emerging to cause relapses. This biological feature makes this parasite resilient to conventional control measures that are designed to treat *Plasmodium falciparum*. Hypnozoites can be treated with primaquine or tafenoquine, but there are contraindications for people with the glucose-6-phosphate dehydrogenase deficiency or for pregnant women and young children (WHO, 2015). Additionally, there are indications that chloroquine resistance in *P. vivax* has emerged in areas of the GMS, which may further create a challenge for treatment (Price et al., 2014). There is no working *in vitro* system to culture *P. vivax* and therefore researchers are increasingly using whole genome studies from field samples in order to answer questions about *P. vivax* evolution (Winter et al., 2015; Hupalo et al., 2016; Pearson et al., 2016; Chen et al., 2017; de Oliveira et al., 2017). However, the lack of samples with high parasitemia has also created barriers to creating necessary genomic resources such as reference genomes.

After the completion of the *P. falciparum* 3D7 genome there was ambition to build a similar genomic reference for *P. vivax* (Carlton, 2003). However, the inability to continuously culture *P. vivax in vitro* poses a barrier to using the same methods (Carlton et al., 2008). The original assembly of the Sal-I strain from El Salvador was highly fragmented consisting of over 2000 extra-chromosomal contigs, and as technology improved, it was found to be missing over 800 genes (Hester et al., 2013). In recent years the focus has been on assembling *P. vivax* genomes based on short reads generated from Illumina sequencing paired with bioinformatic processing through increasingly advanced assembly algorithms. Advancing technology has provided a closer view of the core *P. vivax* genome content while gaining additional insights into the nuances of variability within genomes (Hester et al., 2013; Auburn et al., 2016; Chen et al., 2017). In particular, the assembly of the PvP01 strain from Papua Indonesia revealed an astoundingly large repertoire of subtelomeric multi-gene families (Auburn et al., 2016). It has become clear that there is a high degree of genetic heterogeneity in *P. vivax* parasites (Hupalo et al., 2016). Therefore, researchers have begun to build draft *P. vivax* genomes corresponding to different geographical regions (Menard et al., 2013; Auburn et al., 2016; Chen et al., 2017). Previously, we have shown that malaria parasites from the China-Myanmar border were genetically separated from other parts of the GMS and potentially represent a unique genetic profile resulting from the clonal expansion of a few parasite strains during an outbreak (Brashear et al., 2020). Unfortunately, the closest *de novo* assembly and analysis performed previously from this region was heavily fragmented, consisting of over 8,000 different scaffolds and identifying a small portion of subtelomeric *pir* genes (Chen et al., 2017). Thus, we generated representative *P. vivax* genomes from this region to provide additional insights into the evolution of this parasite population. Here, we performed *de novo* genome assembly on whole genome sequencing (WGS) data from four *P. vivax* field isolates from the China-Myanmar border.

MATERIALS AND METHODS

Ethics Statement and Sample Collection

This research was approved by local Bureau of Health of Kachin and the Institutional Review Board of Pennsylvania State University IRB #34319. Patients with acute *P. vivax* malaria presented at the hospital and clinics of Laiza town, Kachin State, Myanmar and Nabang Township, Yunnan Province, China were recruited after obtaining written informed consent. Malaria diagnosis was performed by microscopic examination of Giemsa-stained blood smears. Five milliliters of venous blood were drawn by venipuncture from each patient and filtered to remove human leukocytes as previously described (Li et al., 2017). Parasites were released after lysis of the red blood cells (RBCs) with saponin, pelleted by centrifugation, and stored at −80°C until DNA extraction.

DNA Library Preparation and Illumina Sequencing

Parasite DNA was extracted from the parasite pellets using the QIAamp DNA Mini Kit (Qiagen, Hilde, Germany) and air-dried for storage. DNA from each sample was resuspended in 20 μL of water and DNA concentration was measured on a Qubit Fluorometer. DNA libraries were prepared using the TruSeq Nano DNA Library Prep Kit with up to 100 ng of DNA, which was sheared to create 350 nt inserts. DNA libraries were sequenced on the HiSeq 2500 in rapid mode to create 150 bp paired-end reads. DNA reads were trimmed for quality, removing areas with a Phred score under 20 using trimmomatic (Bolger et al., 2014).

Selection of Isolates for Genome Assembly

Reads were aligned to the reference PvP01 genome using BWA MEM (Li and Durbin, 2010). Samtools v1.3 was used to characterize average coverage over the reference genome, and those with an average coverage under 10 × were excluded (Li et al., 2009). To account for multiple infections, we estimated multiplicity using estMOI (Assefa et al., 2014). For confirmation, we cross referenced these results with visual analysis of the highly polymorphic merozoite surface protein 1 (*MSP1*) gene using integrative genomics viewer (IGV) (Thorvaldsdottir et al., 2013). We excluded samples for which multiplicity was higher than 1. For samples that passed these filters, *P. vivax* genomes were assembled using Velvet assembler at k-mer of 31, 51, 71 and 91 (Zerbino and Birney, 2008). Assembly quality statistics were assessed via QUAST (Gurevich et al., 2013) with manual review focusing on N50, genome length, and the proportion of misassembled contigs. From these statistics we chose the best-performing k-mer for each sample; if two assemblies seemed comparable, we chose the larger of the two in order to better resolve repetitive regions. To improve the assembly and reduce

the number of contigs, each sample's best assembly was scaffolded with SSPACE (Boetzer et al., 2011), had gaps filled with gapfiller (Nadalin et al., 2012), and was error-corrected with Pilon (Walker et al., 2014).

Genome Refinement

The four samples performed the best in terms of N50 and contiguity were pseudo-contiguated with ABACAS2 (Assefa et al., 2009), gap-filled with IMAGE (Tsai et al., 2010) and corrected with Pilon. Assembled and corrected genomes were uploaded to the Companion webserver for multi-faceted gene annotation (Steinbiss et al., 2016). PvP01, obtained from a patient from Papua Indonesia, was used as the reference genome (Auburn et al., 2016), and annotations were performed with reference protein evidence as well as RATT annotation transfer at the strain level. All other options were left at default options, including pseudo-contiguation with a 500 bp match length and 85% similarity. For consisteny between genome annotation, assembled genome statistics for new assemblies including size, N50 and #Ns/kbp were calculated using post-companion genomes using QUAST, omitting sequences less than 500 bp. Because companion pseudo-contiguation bins extrachromosomal scaffolds together and to minimize chimeric extrachromosomal contigs, we separated each stretch of nucleotides separated by 100 or more N characters.

Gene Characterization

GFF files were downloaded for each sample from the companion annotation and used for further gene analysis. Gene lists and predicted protein sequences were extracted from the genome of each field sample and the PvP01 reference genome. Genes were either identified by the PvP01 orthologs or their Pfam identifiers when possible. Additionally, genes were highlighted along each chromosome using Circos (Krzywinski et al., 2009). In addition, Circos plots visualized the depth of supporting reads based on BWA alignment and segments without high identity to P01 which were found using blat comparisons using a 98% identity threshold and the resulting UCSC chain files were used to identify conserved regions and the remaining regions were plotted (Kent, 2002).

Gene predictions from each assembly were input into OrthoVenn2 (Xu et al., 2019) to detect overall ortholog conservation between assemblies. For the MCL analysis, we used an e-value of 1×10^{-5} and an inflation factor of 1. The top four with the lowest number of gaps were used to characterize the gene content of each assembly. Reference genes were annotated using PlasmoDB[1], while new genes were annotated using a combination of companion annotations and BLAST search of PlasmoDB and GenBank.

A combination of PFAM-based and custom Profile Hidden Markov Models (HMMs) were used to detect the presence of gene families. HMMer was then used to scan the predicted genes of each genome assembly (Eddy, 2011). For confirmation, HMM searches were performed on both the Sal-I and PvP01 reference genomes. Additional protein family profiles were constructed for each of 10 *pir* gene subfamilies based on previously classified PIR proteins (Apweiler et al., 2004). Skylign was used for visualization of HMM profile motifs, except for the DUF3671 group which was selected from the PFAM database (Wheeler et al., 2014).

[1] www.plasmodb.org

Phylogeny and Genetic Diversity

Reads for PvP01 and each of the four samples were aligned to the PvP01 reference using BWA-MEM. We then created a maximum likelihood tree with RAXML using the GTR model with Lewis ascertainment correction. In all cases bootstrapping was set to test for convergence (bootstopping) with a maximum of 1000 iterations, and in the case of the whole genome phylogeny, convergence was reached after 50 bootstrapping replicates. The phylogenetic trees were visualized using ggtree (Yu et al., 2017). Similarly, phylogenetic relationships of *pir* and *msp7* genes in four newly assembled *P. vivax* genomes were evaluated using the maximum likelihood algorithm in RAxML-NG with a JTT model. Additionally, we performed identical variant calling against the PvP01 reference for the four assembled samples and analyzed their nucleotide diversity in 1 kb sliding windows using vcftools (Danecek et al., 2011).

Copy Number Variation (CNV) Detection

Copy number variation for four new genome assemblies were identified as previously with several modifications (Huckaby et al., 2019). Briefly, BWA-MEM was utilized to align reads with default settings to the PvP01 reference genome. Breakpoints of the CNVs were identified using the Speedseq pipeline which includes CNVnator for read-depth analysis and LUMPY for discordant and split-reads analysis. Additionally, a final Speedseq step included a Bayesian likelihood analysis for genotyping of structural variants, which analyzes the ratio and quantity of discordant or split-reads to concordant read-pairs and reads. The settings were modified to require a minimum of two supporting discordant read pairs or split-reads for LUMPY and utilize 1000 bp windows for CNVnator read-depth analysis and the option for genotyping by SVTyper within Speedseq (Chiang et al., 2015).

To generate a final high confidence CNV set, several exclusion criteria were applied. First, CNVs were excluded if their lengths were <2000 bp or >100,000 bp. CNVnator with 1000 bp windows cannot call CNVs <2000 bp and CNVs > 100,000 bp are frequently false positives. Next, CNVs with Speedseq quality scores of <0 were disregarded in order to exclude randomly mapped discordant read-pairs and split-reads (due to sequence homology at multiple locations). CNVs on PvP01 contigs were also excluded as the confidence in their assembly is less than the full PvP01 chromosomes. For final high-confidence CNV calls, any CNVs without > 50% reciprocal overlap of both LUMPY and CNVnator calls were excluded. The boundaries of the resulting high-quality CNVs were reported using the LUMPY locations as discordant read-pairs and split-reads give greater resolution for breakpoints. When multiple LUMPY calls were found to overlap in a single sample, and these resulted in a similar length to a CNVnator call (i.e., CNVs on chromosome 9), boundaries were reported as the outer limits of the LUMPY CNV calls.

RESULTS

Genome Sequencing and Assembly

We obtained 23 *P. vivax* field samples from the China-Myanmar border: four from the Nabang town on the China side and 19 from Laiza town on the Myanmar side. The number of reads obtained for each sample ranged from 7,252,842 to 9,529,810 with the median being 8,387,821. After removing poor-quality reads and reads matching the human genome, the remaining number of reads ranged from 376,875 (4.3% of total reads) to 6,663,732 (77.9%) with the median being 3,278,840 (36.7%). Alignment to the PvP01 reference genome revealed that the average depth per nucleotide for each sample ranged from 2.6 to 61.7 with the median average depth being 29.1 (**Supplementary Figure S1A**).

We performed *de novo* assembly on 14 field samples of *P. vivax* infections with the highest coverage, and the assembled genomes had a median N50 of 38,858.5 and a median number of contigs of 2143 (**Supplementary Figures S1B–D**). Four samples with the highest N50 values prior to correction, and confirmed as monoclonal using three separate methods (**Supplementary Figure S2**), were selected for detailed characterization. These were pseudo-contiguated and corrected for downstream analysis. The median depth of these four samples was 58.87 and coverage remained consistently high throughout the chromosomes (**Figure 1**). After processing, the four *P. vivax* genomes had a median N50 of 1,781,807, a median number of scaffolds of 83.5, and average genome size of 29.86 Mbp, slightly larger than the PvP01 reference (**Table 1**). After excluding the extra-chromosomal contigs, apicoplast and mitochondrial genomes, on average, 92.80% of chromosomes of the four genomes aligned to PvP01.

Genetic Diversity

Our previous analyses of single nucleotide polymorphisms (SNPs) from samples collected at the China-Myanmar border revealed that three samples (LZCH1720, LZCH1476, and LZCH1886) were closely related (Brashear et al., 2020). To estimate diversity of the four genomes used for assembly, we characterized the degree of diversity within the primary chromosomes. We identified 110,635 high-quality bi-allelic SNPs compared to the PvP01 reference. Sliding window analysis of nucleotide diversity revealed high diversity at the subtelomeric regions and the median nucleotide diversity among all windows was 4.29×10^{-4} (**Supplementary Figure S3**).

Genome Annotations

Genes were predicted for each of the four newly assembled *P. vivax* genomes. On average 6721 genes were predicted as compared to 6741 genes for the PvP01 genome. To validate the post-assembly pipeline, the PvP01 genomes were put through an identical post-processing and gene finding pipeline. The results showed 99.85% identity for the PvP01 gene annotation within the core chromosomes. This high degree of consistency allowed us to use polypeptides from the published PvP01 annotation as the reference for comparison with the new assemblies. Meanwhile, new assemblies consistently had between 92.5 and 93% of their chromosomes aligned to those of the PvP01, whereas the Sal-I assembly only had 90.75% of its chromosomes aligned to those of the PvP01 reference. In each of the four assemblies, areas of low conservation were concentrated around the chromosome ends (**Figure 1**).

The predicted genes for the four new assemblies as well as PvP01 and Sal-I references were assigned into orthologous groups, which resulted in 6916 orthology clusters including 4843 single-copy gene clusters (**Figure 2**). The four genomes had an average of 6598 predicted peptides for comparison to the Sal-I genome (5389 annotations) and PvP01 genome (6677 annotations) (**Table 2**). There were 514 coding genes which were present in reference genomes but missing in all four new assemblies (**Figure 2**). These genes primarily consisted of multi-gene family members including *pir* (336), *phist* (4), *stp1* (2), and *msp3* (3). Additionally, many of the missing genes encode proteins with unknown functions such as exported proteins with unknown functions (138), conserved *Plasmodium* proteins with unknown function (9) and hypothetical proteins (9). Additional genes annotated in the reference genomes but not found in the new assemblies include 4 apicoplast ribosomal proteins, 2 ribosomal proteins, 1 reticulocyte binding protein 1a, 1 putative helicase, 1 carbonic anhydrase pseudogene, 1 PPPDE peptidase, 1 palmitoyltransferase DHHC9, 1 porphobilinogen synthase, and 1 KS1 protein precursor.

Newly Identified Genes

There were 151 new ortholog clusters which contained genes without direct orthologs in PvP01; 120 of these contained just one gene from each of the four assemblies. Of the 120 new clusters, 65 were characterized as *pir* genes, 22 were proteins of the unknown functions and/or had the domain of unknown function, DUF3671, and 33 were hypothetical proteins with no known pfam domain match. The 31 gene clusters which had more than one ortholog in at least one genome assembly accounted for 44 new genes in NB45, 45 new genes in LZCH1720, and 46 new genes in both LZCH1476 and LZCH1886. Of the 31 genes, 19 were *pir* genes, six were genes of unknown function, and six were hypothetical proteins with no known pfam domain match.

Multigene Families

The majority of differences noted were within multigene family members, and areas of low conservation seemed to coincide with subtelomeric regions while the core genomes were relatively conserved (**Figure 1**). This creates the possibility that many new or missing genes may be mutually present as part of highly variable gene families, and are therefore very divergent from their P01 orthologs. To verify the overall genome assembly results of the multigene families, we extracted all members for the *pir*, *phist*, *stp1*, *rbp*, *msp7,* and *etramp* gene families from the four new *P. vivax* assemblies and the reference genomes Sal-I and PvP01 (**Table 2**). The *pir* gene family repertoire ranged from 1111 to 1204 in the four new *P. vivax* genomes as compared to 1181 in the PvP01 genome and 387 in the Sal-I genome. Three new assemblies had 8–10 *stp1* family members similar to PvP01 with 10 members, whereas the NB45 genome had 16. Each of the four newly assembled *P. vivax* genomes had 76 or 77 *phist*

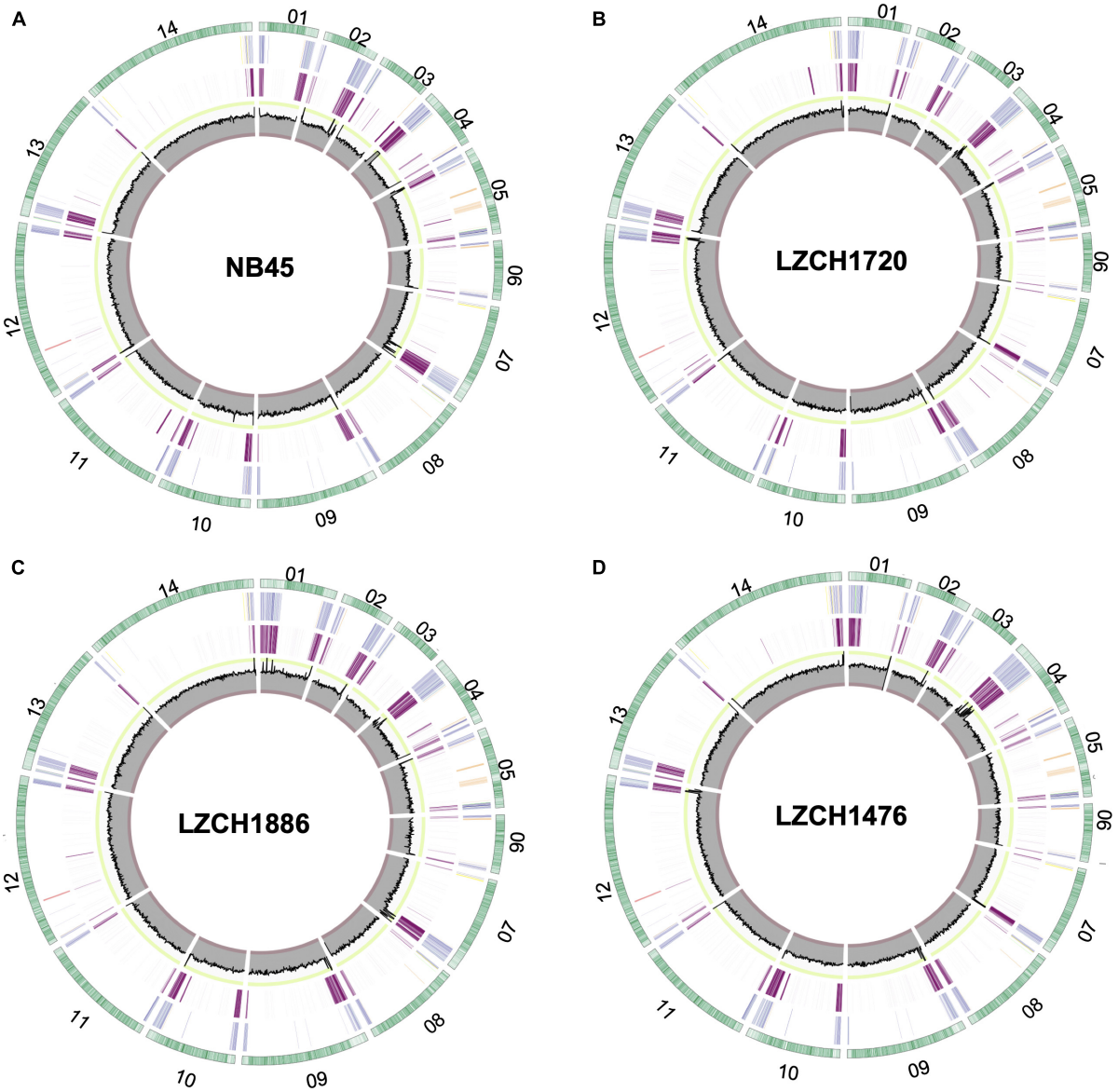


FIGURE 1 | Circos plots for the four *P. vivax* genome assemblies **(A)** NB45, **(B)** LZCH1720, **(C)** LZCH1886, and **(D)** LZCH1476. The outermost green circle represents genes on each chromosome. The second ring colors in select gene families wherein PIR genes are blue, DUF3671 genes are teal, MSP7 genes are red, RBPs are yellow, PHIST genes are orange, and STP1 genes are green. The 3rd ring highlights regions which do not share 98% identity with PvP01. The innermost circle represents coverage of reads for the assembly wherein the minimum is 0, the maximum is 100. To provide context, the interval between 90 and 100 is shown in green, and the interval between 0 and 10 is shown in red.

family members (compared to 79 and 81 in Sal-I and PvP01, respectively), 9 *rbp* genes (compared to 10 in both references), and 40 *tryptophan-rich antigen* genes (compared to 36 and 40 in Sal-I and PvP01, respectively). Each of the new assemblies had the same number of genes as in Sal-I and PvP01 for *msp7* (11 members) and *etramp* (10 members). It is noteworthy that due to high levels of diversity, some of the gene family members were not clustered via orthology with genes in PvP01. For example, of the 548 genes without a PvP01 ortholog in LZCH1476, 356 were *pir* genes, 77 had similarity to DUF3671 (a gene with an unknown function), and one was an *stp1* gene. Overall, of the genes within new multi-gene orthology clusters, the majority (between 30 and 32 genes in each sample) had domain similarity to *pir* proteins, and between 7 and 11 for each assembly had DUF3671 domain structures. Segments containing many of gene family members, particularly *pir* and *msp7* gene family members, coincided with low conservation regions (**Figure 1**).

Pir Subfamily Identification

To establish that missing orthology within a subset of *pir* genes was not an outcome of missing gene family members but rather due to more stringent classification methods, we assigned the

TABLE 1 | Genome statistics for each of the four *P. vivax* field samples and two reference genomes.

	Sal-I	P01	NB45	LZCH1476	LZCH1886	LZCH1720
Reads sequenced	NA	NA	15343644	12553940	12799214	12929808
Average depth			58.51	58.04	59.23	60.07
Total genome length (Mbp)	27.01	29.05	30.05	29.88	29.53	29.84
Chromosome length (Mbp)	22.62	24.21	25.00	25.54	25.61	24.93
Ns/100 kbp	199.53	517.08	1034	1327.3	1365.97	1337.33
Number of scaffolds	2748	242	139	73	65	94
N50 (bp)	1678596	1761288	1743222	1783114	1780500	1920463
% Alignment to P01	79.31%	99.48%	81.43%	81.50%	81.06%	81.36%

extracted *pir* genes to *pir* gene subfamilies as previously defined (Lopez et al., 2013). In each of the four newly assembled samples, nearly all *pir* genes were mapped to at least one *PIR* subfamily (**Table 3**). In the LZCH1886, LZCH1476, LZCH1720, and NB45 assemblies, there were 35, 38, 38 and 33 predicted PIR proteins, respectively, unassigned to PIR subfamilies. Because our reported estimates of *pir* genes is similar to previously reported (Auburn et al., 2016), and in order to establish subfamily diversity, we examined the composition of *pir* subfamilies. We created a maximum likelihood tree for predicted PIR proteins in two selected samples. We found clustering of samples from the same gene subfamilies, with the largest gene subfamily consistently being PIR E, closely followed by PIR C. Interestingly, while the majority of unassigned PIR proteins formed a clade with the PIR C subfamily in the NB45 genome assembly, they formed a more distinct clade in the LZCH1720 assembly (**Figure 3**).

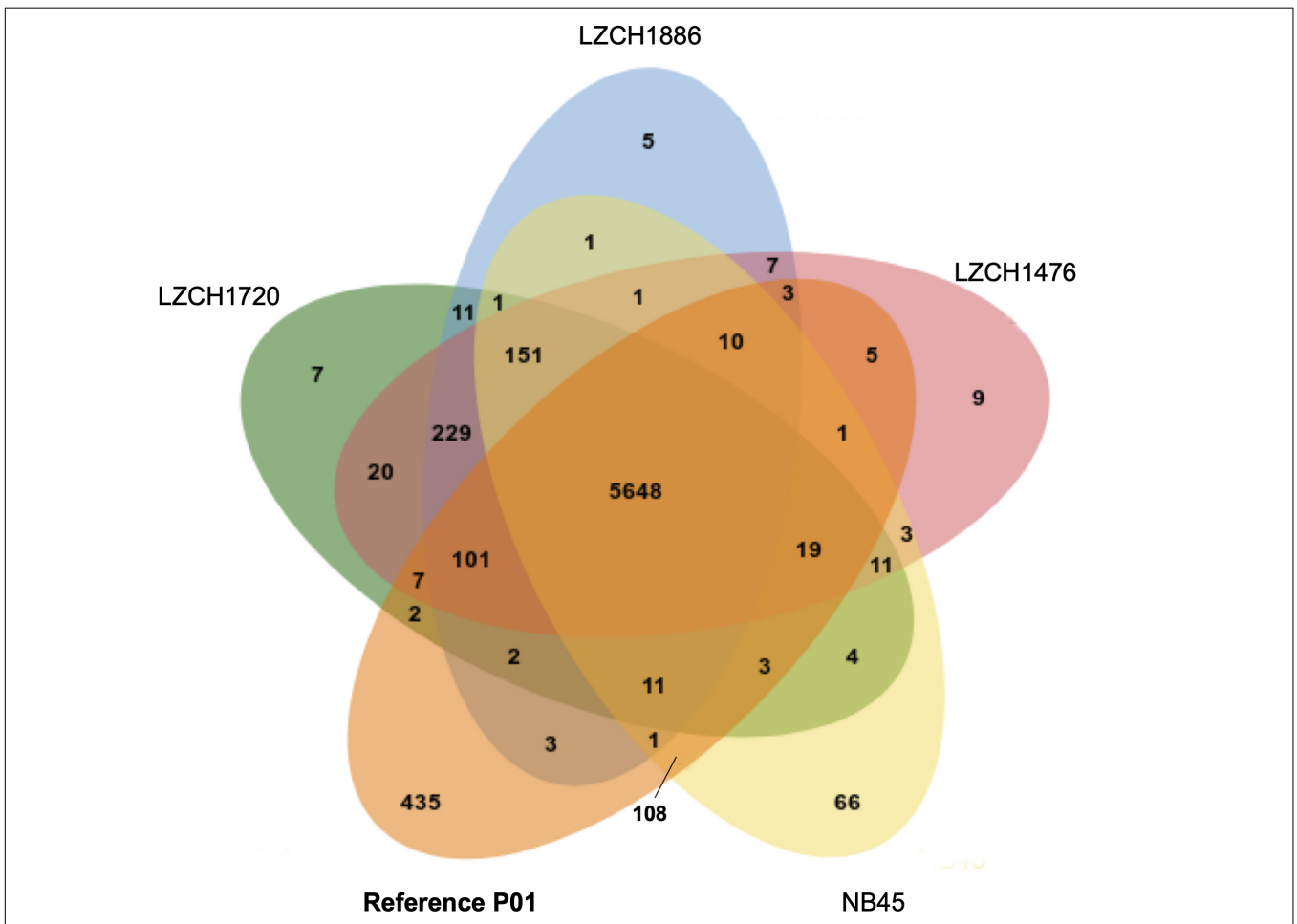


FIGURE 2 | Ortholog presence in each of the four China-Myanmar border *P. vivax* isolates compared to reference PvP01 genome. Each color represents all ortholog groups present in each of 5 different isolates, with the reference P01 assembly being shown in orange.

TABLE 2 | Gene content of each of 4 genomes chosen for gene analysis compared to references Sal-I and PvP01.

Annotation	Sal-I	P01*	LZCH1720	LZCH1886	NB45	LZCH1476
Encoded peptides	5389	6677	6606	6530	6655	6601
Genes without PvP01 orthologs	60	34*	712	543	869	548
PIR	387	1181	1157	1111	1204	1164
ETRAMP	10	10	10	10	10	10
PHIST	79	81	77	76	77	76
STP1*	13	10	8	10	16	9
RBP*	10	10	9	9	9	9
MSP7	11	11	11	11	11	11
Tryptophan-rich antigens	36	40	40	40	40	40

**P01 numbers reflect an identical annotation pipeline imposed on the P01 genome sequence in order to control for differences in methodology and are therefore not identical to existing annotations.*

TABLE 3 | Count of PIR protein subfamilies within PIR orthologous genes.

Subfamilies	LZCH1720	LZCH1886	NB45	LZCH1476
PIR A	32	32	37	32
PIR B	53	47	51	53
PIR C	227	217	225	227
PIR D	0	0	0	0
PIR E	231	225	256	238
PIR G	117	109	130	114
PIR H	51	47	47	51
PIR I	124	122	128	124
PIR J	138	133	143	138
PIR K	154	144	154	154
Unassigned	38	35	33	38

The subfamily D had no members within the PIR families of any of the newly assembled *P. vivax* genomes (**Table 3**). To establish what may be causing a depletion of this gene family, we ran an HMM profile scan against all proteins in the genome using previously identified PIR D family members and found that there were 270, 280, 260 and 272 proteins within the entire predicted protein database of the newly assembled *P. vivax* genomes LZCH1720, NB45, LZCH1886, and LZCH1476, respectively. Each of the identified proteins had high similarity scores to the profile of the constructed PIR D subfamily. Interestingly, despite these similarities to the PIR D subfamily, most of the proteins had higher similarity to the DUF3671 domain. Specifically, PIR D subfamily members accounted for 253/254 (99.6%), 237/238 (99.5%), 247/248 (99.6%), and 259/260 (99.6%) of DUF3671 proteins in LZCH1476, LZCH1886, LZCH1720 and NB45, respectively.

To determine if DUF3671 domain characterizes a separate gene family which overlaps with the previously designated PIR D gene subfamily, we performed analysis of both gene domain homologies. Within the Pfam database there are 489 sequences which contain the DUF3671 domain, all of which belong to a species of non-falciparum *Plasmodium*; 72% from *P. malariae*. We also compared motifs between the DUF3671 profile available from Pfam, and the constructed HMM profile based on suggested PIR D family members, finding that two of the most conserved profile components from the PIR D profile were similar to two of the most conserved profile components from the DUF3671 profile (**Supplementary Figure S4**). Meanwhile, neither of the architecture components were present in the overall PIR protein profile, and none of the major components of the PIR protein profile were present in the profile of PIR D.

MSP7 Diversity

Genes from the *msp7* family were consistently represented with 11 genes in each assembly (**Table 2**). *Msp7* genes tend to be variable and are therefore not well characterized by SNP data. Our assemblies all showed low conservation to PvP01 around the *msp7* locus (**Figure 1**), so we characterized diversity within the *msp7* gene family within the newly assembled genomes (**Figure 4**, **Supplementary Figure S5**). To establish general relatedness of the parasite for comparison, we first constructed a phylogenetic tree based on SNPs from the four new assemblies (**Figure 4A**, **Supplementary Figure S5B**), which consistently revealed the closer relationship among three new genome assemblies. Phylogenetic tree of predicted MSP7 proteins showed that the MSP7-C and MSP7-A orthologs, consisting of two protein clades each, were highly related between these assemblies (**Figure 4B**, **Supplementary Figure S5**), highlighting evolutionary conservation of these genes. MSP7-B subfamilies, consisting of the remaining 7 proteins, were much more diverse. Within each protein cluster, the PvP01 and Sal-I MSP7-B protein formed a separate clade from the MSP7-B proteins from China-Myanmar border assemblies (**Figure 4B**, **Supplementary Figure S5B**). In two instances, however, NB45 appeared to form a clade with the PvP01 and Sal-I references rather than the other China-Myanmar border samples (**Figure 4B**, **Supplementary Figure S5B**).

Copy Number Variations (CNVs)

Copy Number Variations in the new genome assemblies were scanned using different algorithms. After applying several criteria to identify high-confidence CNVs, we identified a final set of five amplifications (averaging 46 kb) and 11 deletions (averaging 27 kb) in the four new assemblies. All of the amplifications were supported by more than one sample and were located in telomeric or subtelomeric regions (**Table 4**, **Figure 5A**). A set of three amplifications found in LZCH1720 and NB45 on chromosome 9 appeared to exhibit complex breakpoints as two of the CNVs that were called added up to the largest CNV found at that location (**Table 4**). Of the deletions, five were identified in individual samples and the other six were present in two or three samples (**Table 4**, **Figure 5**). A similar situation was found on chromosome 9 with multiple deletions found just upstream of the amplification loci that were previously mentioned (**Table 4**, **Figure 5**). This locus was found to be variable in all four assemblies, with LZCH1476 and LZCH1720 having the same deletion (**Table 4**, **Figure 5A**).

Based on current annotations for the PvP01 genome, the majority of genes included in the CNV amplifications were *pir*

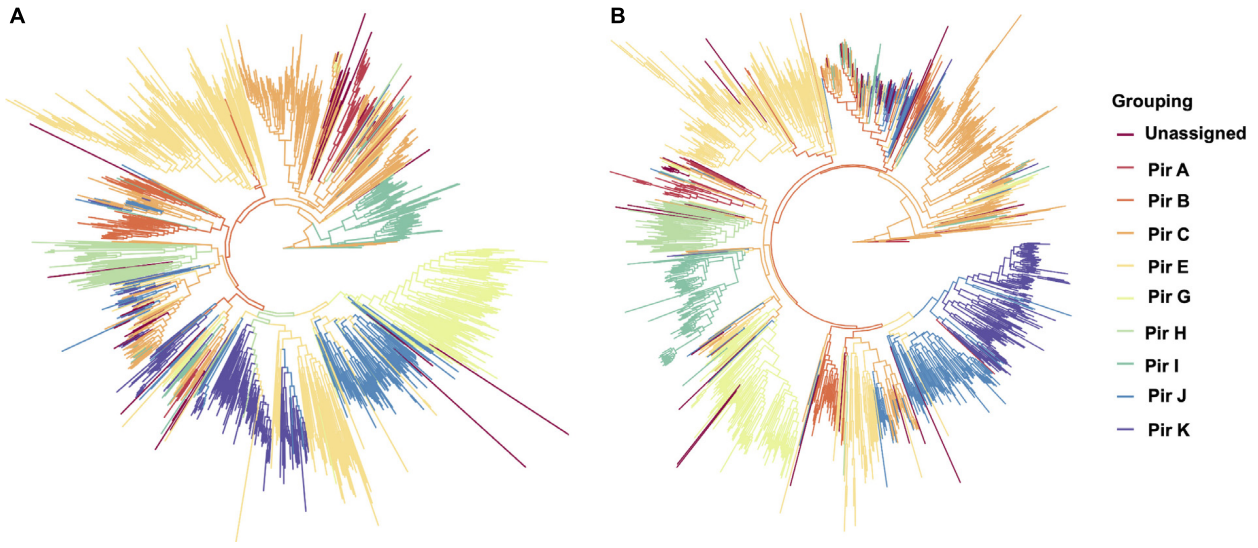


FIGURE 3 | Maximum-likelihood PIR family structure within two assembled genomes **(A)** NB45 and **(B)** LZCH1720. Each color represents a separate subfamily as denoted by the color key on the right.

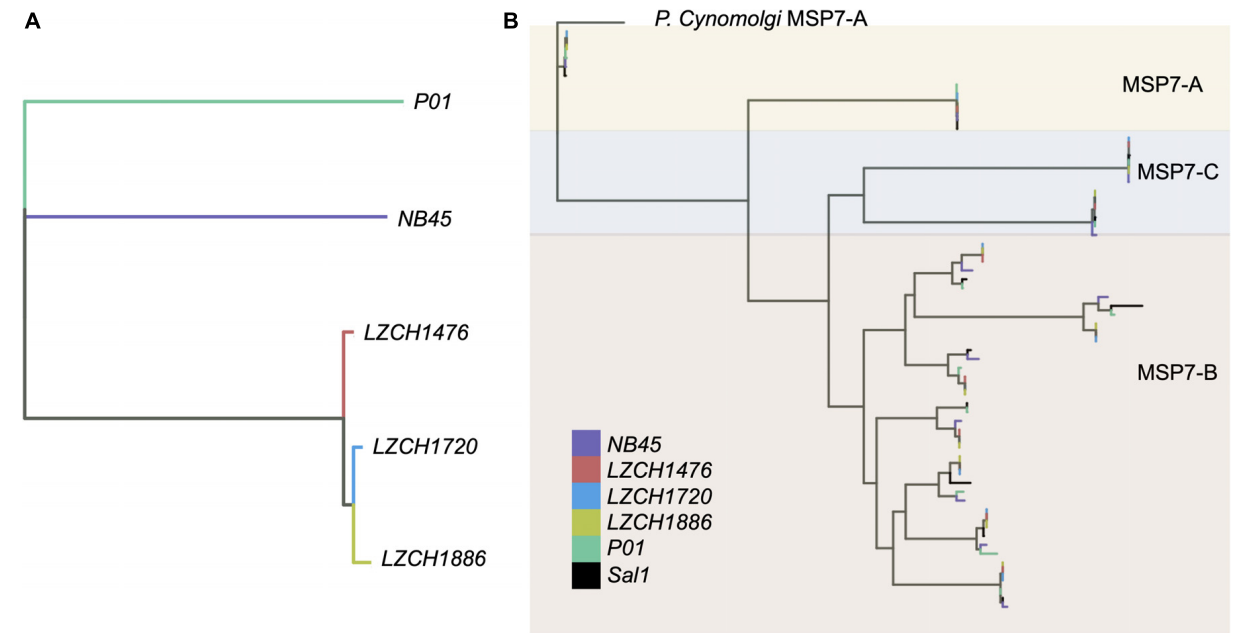


FIGURE 4 | Genetic diversity within *P. vivax* assemblies. **(A)** Genetic diversity within the entire genome of the four assemblies compared to PvP01 based on SNPs from alignment to Sal-I. **(B)** Diversity within *msp7* genes between different assemblies, Sal-I and PvP01. *P. cynomolgi* MSP7-A was used as an outgroup. Colors: Black– P01, Blue – Sal-I, Orange—NB45, Red—LZCH1720, Cyan—LZCH1476, Purple—LZCH1886.

family members which matches results from the assemblies. However, the amplifications on chromosomes 4 and 14 also included putative exported proteins with unknown functions (**Supplementary File 1**). The genes involved in deletions were similar to the amplifications with the majority composed of *pir* family members or exported proteins with unknown functions (**Supplementary File 1**). The complex amplifications and deletions on chromosome 9 included hypothetical proteins.

Genes within the predicted CNVs in the genome assemblies involve many hypervariable gene families. Despite lower resolution of the hyper-variable regions within the newly assembled genomes, direct comparisons to the assembled genome in some cases further corroborated the detection of CNVs. Within chromosome 2, for instance, deletions were predicted on either end of the chromosome for LZCH1720, LZCH1886 and LZCH1476, but not in NB45 (**Figure 5**, **Table 4**).

TABLE 4 | Copy Number Variations detected within the 4 isolates compared to PvP01.

Supporting Samples	Type	Chr.	Start	End	Length
LZCH1886, NB45	Amplification	4	38123	124716	86593
LZCH1476, LZCH1720, NB45	Amplification	9	79721	135397	55676
LZCH1720, NB45	Amplification	9	79721	94710	14989
LZCH1720, NB45	Amplification	9	94119	135206	41087
LZCH1476, LZCH1720	Amplification	14	82	35455	35373
NB45	Deletion	1	58587	62858	4271
LZCH1476, LZCH1720, LZCH1886	Deletion	2	15890	19643	3753
LZCH1476, LZCH1720, LZCH1886	Deletion	2	776331	785302	8971
LZCH1720	Deletion	5	1010209	1013244	3035
LZCH1720, LZCH1886	Deletion	7	1531179	1630181	99002
LZCH1720	Deletion	8	1647230	1662256	15026
LZCH1476, LZCH1720	Deletion	9	135520	190955	55435
LZCH1886	Deletion	9	135520	182436	46916
NB45	Deletion	9	141424	191395	49971
LZCH1476, LZCH1720, LZCH1886	Deletion	10	91532	96146	4614
LZCH1720, LZCH1886	Deletion	11	2054089	2057879	3790

Chr., chromosome.

Accordingly, while NB45 had regions homologous to the deleted region within its assembled genome, the other three assemblies were missing this section (**Figure 5B**). Visualization of one of the two regions, between 15890 and 19643 further supported a reduction in properly paired read coverage and increases of surrounding discordant reads for the predicted region for all assemblies except for NB45.

DISCUSSION

To date, no gapless *P. vivax* genome has been completed, and diversity within *P. vivax* populations results in genomic differences for populations from different geographic areas. The China-Myanmar border, for instance, represents a unique region of malaria endemicity, which would benefit from the availability of high-quality *P. vivax* genome sequence collection from this area. Here we present four draft genomes with similar contiguity to the published reference PvP01 and performed detailed gene annotations. We have demonstrated that their genomic contents are consistent with expectations for the core *P. vivax* genomes. However, some genes, particularly those from antigenic gene families, are highly diverse and frequently not directly orthologous to their counterparts in the current reference genomes.

The *P. vivax* genome assemblies described here are of comparable or higher quality to the existing references. In terms of length and contiguity they parallel the genome assemblies from Papua Indonesia, China and Thailand, which ranged between 28.9 and 30.2 Mbp with 230–529 scaffolds, compared to our samples which covered between 28.5 and 30.1 Mbp within 65–139 scaffolds (Auburn et al., 2016; **Table 1**). Between 92.5% and 93% of the 14 chromosomes, representing the core nuclear genome, from each assembly aligned to the reference PvP01. Missing bases on the genome were common, ranging from 1034 to 1365.97 Ns/Kbp compared to 517 in the reference. However, we were still able to capture the majority of annotated genes, comparable to or even in excess of the PvP01 reference (**Table 2**). Only six genes missing from our assemblies were not a member of a multi-gene family. It is worth noting that PvP01 is the most contiguous and highest-coverage assembly used currently, with other assemblies being much more fragmented. It is noteworthy that both the current gold-standard reference and the new assemblies presented here contain a number of ambiguous sequences and extra-chromosomal contigs, which are intrinsic to the short-read sequencing technology. The next major improvements in genome contiguity will likely occur when *P. vivax* genomic DNA can be obtained in the quantities required for long-read sequencing.

Our data indicate few large-scale variations in the core genome and high variability in multi-gene families and subtelomeric regions, also supported by high nucleotide diversity and occurrence of CNVs toward chromosomal ends (**Supplementary Figure S1**, **Figure 5**). Subtelomeric regions tend to be subject to more frequent genomic rearrangements in malaria parasite, while the homology of the multi-gene families also reduces the accuracy of read mapping technologies. Our recent study found high coverage variability within chromosomal ends (Brashear et al., 2020), and upticks of coverage identified here toward chromosomal ends could be a result of read mismapping (**Figure 1**).

Despite seemingly close relationships (**Table 2**), proteins from multi-gene families were divergent enough from the reference that orthology-based analysis was unable to capture a large number of them (**Figure 2**), even though we were able to capture similar numbers of gene family members within each new *P. vivax* assembly using HMM profile searches. We were also capable of identifying large-scale structural variation around the *pir* genes and genes for exported proteins using CNV analyses, which suggests substantial genomic rearrangements around these gene families (**Supplementary File 1**). Interestingly, while NB45 is genetically differentiated from the other three genomes and harbors distinct deletions, it also shares some overlap of amplification with other samples, particularly LZCH1720 (**Figures 4A**, **5A**, **Table 4**). Previous analyses have identified that some beneficial structural variations arise independently and are selected in parallel (Nair et al., 2008; Menard et al., 2013; Hostetler et al., 2016). Therefore, future analyses of large-scale CNVs within *P. vivax* could identify similarities. Another interesting observation was amplifications and deletions found in close proximity to one another on chromosome 9, suggesting the existence of chromosomal regions that are prone to genetic deletions or amplifications (**Table 4**).

Pir genes in particular are known for having a large degree of diversity and are divided into subfamilies (Lopez et al., 2013; Auburn et al., 2016; Chen et al., 2017). We analyzed the content

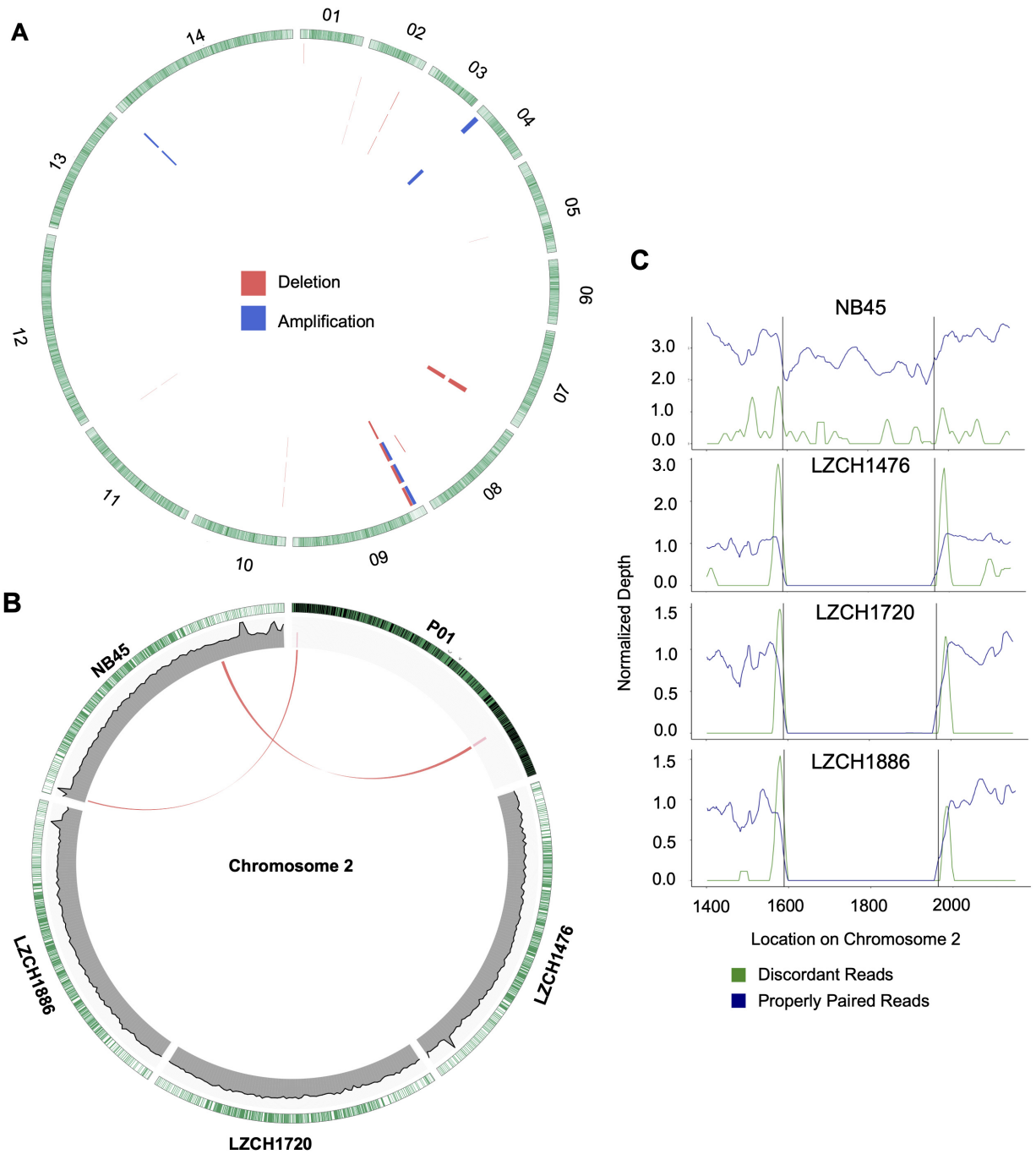


FIGURE 5 | Structural variation within assembled genomes. **(A)** Locations of all predicted copy number variations. Each circle represents a different sample in this order: NB45, LZCH1476, LZCH1720, and LZCH1886. Pink wedges represent predicted deletions while blue wedges represent predicted amplifications. Outer ring shows gene density wherein green highlights are individual genes. **(B)** CNVs on chromosome 2 with wedges representing their location within PvP01 and lines connecting them to homologous regions on respective assemblies when applicable. Coverage support plots were included for the four new assemblies in the second ring. **(C)** Normalized depth for both discordant and properly paired reads around a chromosome 2 deletion predicted for 3 of the 4 samples. Black lines denote the predicted deletion boundary.

of the *pir* gene family obtained through protein profile alignment, and found that subfamilies C and E were the largest gene family members in all assemblies and exhibit a degree of structure even within the subfamilies, which mirrors findings from previous analyses (Auburn et al., 2016; Chen et al., 2017). Interestingly, we found no members of PIR D aligned to the HMM profile for the *pir* gene family, and instead most proteins with high identity to the PIR D subfamily, under isolated searching, had higher

similarity to the DUF3671-containing proteins with unknown functions (**Table 3**). Extensive comparisons of protein families in *P. vivax* has previously demonstrated that PIR subfamilies D, H and A are only loosely related to PIR proteins, with PIR subfamily D being the most dissimilar and having no homology blocks in common with other PIR subfamilies (Lopez et al., 2013). This study posited that PIR A, PIR H and especially PIR D should be considered separate gene families, which our data support. Motif comparisons demonstrated that PIR D subfamily members do not have common motifs with other PIR family members, and instead suggested strong homology with DUF3671. This would explain why previous studies have either represented no gene family members from PIR D, or have specifically used homology to the subfamily grouping to achieve representation (Auburn et al., 2016; Chen et al., 2017). Our data corroborates that PIR D proteins are dissimilar to other *pir* gene family members, and points to an alternative gene family united by the DUF3671 domain. PIR A and PIR H were not as disparate within our dataset and were identifiable by HMM profiling. However, previous phylogenetic analyses comparing various multi-gene families suggest common ancestry between PfEMP1, SicaVAR, and Surfin proteins, making it possible that new gene families could evolve via increasing divergence of existing subfamilies (Frech and Chen, 2013).

We found inter-sample variation within the *msp7*-B genes which would be difficult to detect through SNP profiling alone. Previous studies have shown the diversity of *msp7* within species, finding that *msp7*-A paralogs are well conserved within species (Castillo et al., 2017), a finding well supported here (**Figure 4**). The differences in redundancy and interspecies conservation have previously been suggested to correspond to different roles in invasion for *msp7* subtypes A, B and C (Garzon-Ospina et al., 2016). Patterns of variation within the MSP7-B subfamily closely mimic whole genome findings, suggesting MSP7-B is a viable measure of population diversity (**Figure 4**). NB45 was differentiated from the other three samples in both whole genome analyses and analysis of *msp7* genes (**Figure 4**). Our previous results support that LZCH1720, LZCH1886 and LZCH1476 probably resulted from clonal expansion linked to a *P. vivax* outbreak in 2013, which would explain their higher relatedness as compared to NB45 (Brashear et al., 2020). The small sample size and close relationship among three of the four samples are also likely to result in the relatively low nucleotide diversity as compared to other previously surveyed *P. vivax* populations (Hupalo et al., 2016; de Oliveira et al., 2017).

We present here the first collection of high-quality *P. vivax* genome assemblies from the China-Myanmar border. This area is important in the scheme of malaria elimination due to the potential cross-border spread of malaria parasites (Lo et al., 2015, 2017). Previously, it has been demonstrated that the China-Myanmar border has substantial genetic separation from other areas in Southeast Asia. Further, *P. vivax* is the predominant form of malaria in the border region (Geng et al., 2019). Therefore, in-depth statistics on gene content and variability from *P. vivax* on the China-Myanmar border represent a benefit to the malaria community. Our genome assemblies presented here were able to provide further insight into the dynamics of gene families within *P. vivax* and constitute a new resource for malaria researchers.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI database within bioproject PRJNA603279 (www.ncbi.nlm.nih.gov/bioproject/?term = PRJNA603279).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Bureau of Health of Kachin and the Institutional Review Board of Pennsylvania State University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AB, LC, JG, AH, and JM contributed to study design and conceptualization. QF, YH, YL, YZ, and YC collected blood samples and extracted DNA. AB and ZW prepared genomic sequences. AB prepared assemblies and peptide analysis. AH and LD performed CNV analyses. AB wrote the first draft of the manuscript. All authors contributed to editing, revising, and approving manuscript.

FUNDING

LC received funding from the National Institute for Allergy and Infectious Diseases and the National Institute of Health (NIH) (U19 AI089672). AB received funding from the National Library of Medicine (NIH) (T32 LM012415).

ACKNOWLEDGMENTS

We express gratitude to Drs. Margarita M. López-Uribe, Maciej Boni, Santhosh Giririjan, and Runze Li for their thoughtful comments on experimental design, data analysis, and visualization. We are also grateful to the Penn State Genomics Core Facility for their assistance with DNA sequencing and the University of South Florida Research Computing for use of computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01930/full#supplementary-material

REFERENCES

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi: 10.1093/nar/gkh131

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., and Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 25, 1968–1969. doi: 10.1093/bioinformatics/btp347

Assefa, S. A., Preston, M. D., Campino, S., Ocholla, H., Sutherland, C. J., and Clark, T. G. (2014). estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics* 30, 1292–1294. doi: 10.1093/bioinformatics/btu005

Auburn, S., Bohme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., et al. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res.* 1:4. doi: 10.12688/wellcomeopenres.9876.1

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

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

Brashear, A., Fan, Q., Hu, Y., Li, Y., Zhao, Y., Wang, Z., et al. (2020). Population genomics identifies a distinct *Plasmodium vivax* population on the China-Myanmar border of Southeast Asia. *PLoS Negl. Trop. Dis.* (in press).

Carlton, J. (2003). The *Plasmodium vivax* genome sequencing project. *Trends Parasitol.* 19, 227–231. doi: 10.1016/s1471-4922(03)00066-7

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763. doi: 10.1038/nature07327

Castillo, A. I., Andreina Pacheco, M., and Escalante, A. A. (2017). Evolution of the merozoite surface protein 7 (msp7) family in *Plasmodium vivax* and *P. falciparum*: a comparative approach. *Infect. Genet. Evol.* 50, 7–19. doi: 10.1016/j.meegid.2017.01.024

Chen, S. B., Wang, Y., Kassegne, K., Xu, B., Shen, H. M., and Chen, J. H. (2017). Whole-genome sequencing of a *Plasmodium vivax* clinical isolate exhibits geographical characteristics and high genetic variation in China-Myanmar border area. *BMC Genomics* 18:131. doi: 10.1186/s12864-017-3523-y

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

Cui, L., Yan, G., Sattabongkot, J., Cao, Y., Chen, B., Chen, X., et al. (2012). Malaria in the greater Mekong Subregion: heterogeneity and complexity. *Acta Trop.* 121, 227–239. doi: 10.1016/j.actatropica.2011.02.016

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

de Oliveira, T. C., Rodrigues, P. T., Menezes, M. J., Goncalves-Lopes, R. M., Bastos, M. S., Lima, N. F., et al. (2017). Genome-wide diversity and differentiation in New World populations of the human malaria parasite *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* 11:e0005824. doi: 10.1371/journal.pntd.0005824

Duan, Y., Li, S., Kang, X., Yin, S., and Sun, X. (2013). A point-like outbreak caused by secondary transmission from an imported malaria vivax case. *Int. J. Med. Parasit. Dis.* 40, 57–59.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Frech, C., and Chen, N. (2013). Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *BMC Genomics* 14:427. doi: 10.1186/1471-2164-14-427

Garzon-Ospina, D., Forero-Rodriguez, J., and Patarroyo, M. A. (2016). Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis. *BMC Evol. Biol.* 16:256. doi: 10.1186/s12862-016-0830-x

Geng, J., Malla, P., Zhang, J., Xu, S., Li, C., Zhao, Y., et al. (2019). Increasing trends of malaria in a border area of the Greater Mekong Subregion. *Malar. J.* 18:309. doi: 10.1186/s12936-019-2924-6

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

Hester, J., Chan, E. R., Menard, D., Mercereau-Puijalon, O., Barnwell, J., Zimmerman, P. A., et al. (2013). De novo assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. *PLoS Negl. Trop. Dis.* 7:e2569. doi: 10.1371/journal.pntd.0002569

Hostetler, J. B., Lo, E., Kanjee, U., Amaratunga, C., Suon, S., Sreng, S., et al. (2016). Independent origin and global distribution of distinct *Plasmodium vivax* duffy Binding protein gene duplications. *PLoS Negl. Trop. Dis.* 10:e0005091. doi: 10.1371/journal.pntd.0005091

Huckaby, A. C., Granum, C. S., Carey, M. A., Szlachta, K., Al-Barghouthi, B., Wang, Y. H., et al. (2019). Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome. *Nucleic Acids Res.* 47, 1615–1627. doi: 10.1093/nar/gky1268

Hupalo, D. N., Luo, Z., Melnikov, A., Sutton, P. L., Rogov, P., Escalante, A., et al. (2016). Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat. Genet.* 48, 953–958. doi: 10.1038/ng.3588

Imwong, M., Suwannasin, K., Kunasol, C., Sutawong, K., Mayxay, M., Rekol, H., et al. (2017). The spread of artemisinin-resistant *Plasmodium falciparum* in the Greater Mekong subregion: a molecular epidemiology observational study. *Lancet Infect. Dis.* 17, 491–497. doi: 10.1016/s1473-3099(17)30048-8

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kittichai, V., Koepfli, C., Nguitragool, W., Sattabongkot, J., and Cui, L. (2017). Substantial population structure of *Plasmodium vivax* in Thailand facilitates identification of the sources of residual transmission. *PLoS Negl. Trop. Dis.* 11:e0005930. doi: 10.1371/journal.pntd.0005930

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

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

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

Li, J., Tao, Z., Li, Q., Brashear, A., Wang, Y., Xia, H., et al. (2017). Further evaluation of the NWF filter for the purification of *Plasmodium vivax*-infected erythrocytes. *Malar. J.* 16:201. doi: 10.1186/s12936-017-1855-3

Lo, E., Lam, N., Hemming-Schroeder, E., Nguyen, J., Zhou, G., Lee, M.-C., et al. (2017). Frequent spread of *Plasmodium vivax* malaria maintains high genetic diversity at the Myanmar-China border, without distance and landscape barriers. *J. Infect. Dis.* 216, 1254–1263. doi: 10.1093/infdis/jix106

Lo, E., Zhou, G., Oo, W., Lee, M. C., Baum, E., Felgner, P. L., et al. (2015). Molecular inference of sources and spreading patterns of *Plasmodium falciparum* malaria parasites in internally displaced persons settlements in Myanmar-China border area. *Infect. Genet. Evol.* 33, 189–196. doi: 10.1016/j.meegid.2015.05.002

Lopez, F. J., Bernabeu, M., Fernandez-Becerra, C., and del Portillo, H. A. (2013). A new computational approach redefines the subtelomeric vir superfamily of *Plasmodium vivax*. *BMC Genomics* 14:8. doi: 10.1186/1471-2164-14-8

Menard, D., Chan, E. R., Benedet, C., Ratsimbasoa, A., Kim, S., Chim, P., et al. (2013). Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains. *PLoS Negl. Trop. Dis.* 7:e2489. doi: 10.1371/journal.pntd.0002489

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinform.* 13(Suppl. 14):S8. doi: 10.1186/1471-2105-13-S14-S8

Nair, S., Miller, B., Barends, M., Jaidee, A., Patel, J., Mayxay, M., et al. (2008). Adaptive copy number evolution in malaria parasites. *PLoS Genet.* 4:e1000243. doi: 10.1371/journal.pgen.1000243

Pearson, R. D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C., et al. (2016). Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat. Genet.* 48, 959–964. doi: 10.1038/ng.3599

Price, R. N., von Seidlein, L., Valecha, N., Nosten, F., Baird, J. K., and White, N. J. (2014). Global extent of chloroquine-resistant *Plasmodium vivax*: a systematic

review and meta-analysis. *Lancet Infect. Dis.* 14, 982–991. doi: 10.1016/S1473-3099(14)70855-2

Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., et al. (2016). Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* 44, W29–W34. doi: 10.1093/nar/gkw292

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

Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 11:R41. doi: 10.1186/gb-2010-11-4-r41

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

Wheeler, T. J., Clements, J., and Finn, R. D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinform.* 15:7. doi: 10.1186/1471-2105-15-7

WHO (2015). *Guidelines for the Treatment of Malaria*. Geneva: World Health Organization.

WHO (2016). *Eliminating Malaria in the Greater Mekong Subregion: United to End a Deadly Disease*. Geneva: World Health Organization.

WHO (2018). *World Malaria Report 2018*. Geneva: World Health Organization.

Winter, D. J., Pacheco, M. A., Vallejo, A. F., Schwartz, R. S., Arevalo-Herrera, M., Herrera, S., et al. (2015). Whole genome sequencing of field isolates reveals extensive genetic diversity in *Plasmodium vivax* from Colombia. *PLoS Negl. Trop. Dis.* 9:e0004252. doi: 10.1371/journal.pntd.0004252

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

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

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.

Copyright © 2020 Brashear, Huckaby, Fan, Dillard, Hu, Li, Zhao, Wang, Cao, Miao, Guler and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/fmicb.2020.573430

Comprehensive Genome Analysis of 6,000 USA SARS-CoV-2 Isolates Reveals Haplotype Signatures and Localized Transmission Patterns by State and by Country

*Lishuang Shen[1], Jennifer Dien Bard[1,2], Jaclyn A. Biegel[1,2], Alexander R. Judkins[1,2] and Xiaowu Gai[1,2]**

[1] Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, United States,
[2] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

OPEN ACCESS

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

Reviewed by:
Alan J. Wolfe,
Loyola University Chicago, United States
Hongbo Xie,
Children's Hospital of Philadelphia, United States
Tamim Shaikh,
University of Colorado, United States

***Correspondence:**
Xiaowu Gai
xgai@chla.usc.edu

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

Received: *17 June 2020*
Accepted: *14 August 2020*
Published: *03 September 2020*

Citation:
Shen L, Dien Bard J, Biegel JA, Judkins AR and Gai X (2020) Comprehensive Genome Analysis of 6,000 USA SARS-CoV-2 Isolates Reveals Haplotype Signatures and Localized Transmission Patterns by State and by Country. Front. Microbiol. 11:573430. doi: 10.3389/fmicb.2020.573430

Genomic analysis of SARS-CoV-2 sequences is crucial in determining the effectiveness of prudent safer at home measures in the United States (US). By haplotype analysis of 6,356 US isolates, we identified a pattern of strongly localized outbreaks at the city-, state-, and country-levels, and temporal transmissions. This points to the effectiveness of existing travel restriction policies and public health measures in reducing the transmission of SARS-CoV-2.

Keywords: COVID-19, SARS-CoV-2, haplotype, variant, localized outbreak, temporal transmission

INTRODUCTION

SARS-CoV-2 is a positive-sense single-stranded RNA virus (Wu et al., 2020; Zhu et al., 2020). The collection of variants in a viral genome is referred to as the haplotype. New haplotypes arise from sequential acquisition of new variants in the viral genome. A haplotype, more than individual variants, serves as the distinct signature of a viral isolate and can be used effectively to trace the lineage, determine the ancestral origin of the infection, and to understand the community spread pattern during the COVID-19 pandemic. The rapid accumulation and sharing of SARS-CoV-2 genome sequences at an unprecedented speed have greatly facilitated haplotype-based genomic epidemiology studies. Since the first SARS-CoV-2 genome sequence was reported in January of 2020, as of May 20th 2020, there have been over 30,000 sequences deposited to GISAID (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017),[1] NCBI Virus,[2] the China National Center for Bioinformation (CNCB) 2019 nCoV Resource (Zhao et al., 2020)[3] and other data repositories (Shen et al., 2020b).

To understand the genetic etiology of COVID-19, it is imperative to have a comprehensive understanding of the variant and haplotype landscapes of all reported genomes of SARS-COV-2. Country-, state- and possibly city-specific variant profiles may contribute to varied disease exemplifications and fatality rates observed across the globe along with host factors such as age, ethnicity and comorbidity. In our previous study, we established a comprehensive

[1] https://www.gisaid.org/
[2] https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
[3] https://bigd.big.ac.cn/ncov/

COVID-19 genomic resource, Children's Hospital Los Angeles (CHLA) COVID-19 Analysis Research Database (CARD), by harmonizing data from GISAID, NCBI Virus, CNCB and other resources (Shen et al., 2020b). Leveraging this extensive resource, in this study, we performed a comprehensive study of all publicly available SARS-CoV-2 genome sequences at the time of study, restricted to comparisons of global vs. USA isolates, which included 83 isolated that we sequenced at the Children's Hospital Los Angeles (Shen et al., 2020b). We called variants from each genome sequence, and performed categorical analyses of variants and haplotypes as stratified by the geographic locations. This genomic epidemiology study, focusing on haplotypes, allowed us to gain insights as to the continuous evolution of the SARS-CoV-2 viral genomes and how the travel restrictions and safer-at-home health measures had been effective in reducing the spread of this pandemic by preventing the inter-state transmission of the virus.

MATERIALS AND METHODS

Global and US SARS-CoV-2 Sequence Data

The CHLA internal SARS-CoV-2 sequencing data were generated using the SARS-CoV-2 whole genome sequencing research assay, established by the CHLA Center for Personalized Medicine and the Virology Laboratory. The major external resources of SARS-CoV-2 strains, genome sequences, and variants were GISAID, GenBank, CNCB, and NextStrain. Details about how the sequences were collected and further processed in CHLA COVID-19 Analysis Research Database (CARD) were described previously (Shen et al., 2020b).

Sequence Alignment, Variant Calling, and Haplotype Analysis

Viral genome comparison and variant calling is a wraparound of MUMmer version 4.0.12 (Marçais et al., 2018),[4] with results loaded into a MySQL database. Haplotype analysis was done with SQL queries and custom scripts, all as part of the CHLA COVID-19 Analysis Research Database (CARD) which was described previously (Shen et al., 2020b).

RESULTS

Viral genomes and demographic meta-data of 6,356 SARS-CoV-2 isolates within the US (as of May 20, 2020) were extracted from GISAID,[5] GenBank,[6] and COVID-19 patients and staff at the Children's Hospital Los Angeles (CHLA). Variants, haplotype, geographic location at diagnosis and documented exposure for the patients were analyzed (Shen et al., 2020b). A total of 921 unique variants were each detected in at least three US isolates. Similarly, 264 distinct haplotypes were each represented in at least five US isolates (**Supplementary Table S1**). These variants and haplotypes were hence deemed unlikely to be sequencing artifacts and kept for further analysis.

The four most common mutations (241-C-T, 3037-C-T, 14408-C-T, 23403-A-G) were each present in about 65–67% of US isolates. In total, these 921 variants included 487 missense, 348 synonymous, 66 intergenic, 4 in-frame deletions, 5 stop gained/lost, and several other non-coding variants (**Supplementary Table S2**).

US-Specific Haplotypes

Cross-stratification by geolocation identified city-, state- and country-specific haplotypes. Seventy seven of the 264 (29.2%) haplotypes that were found in at least five US isolates were US-specific. They comprise of a total of 849 isolates which accounted for 13.3% of the 6,356 US isolates. In addition to the 77 purely US-specific haplotypes, there were an additional four large haplotypes that were mostly North America-specific, with a total of 434 isolates where 431 isolates (99.3%) were from the US (425) and Canada (6) (**Supplementary Table S3**). Isolates from these four large haplotypes were geographically dispersed across the nation. Of note, all 66 US isolates belonging to haplotype (241-C-T, 1059-C-T, 3037-C-T, 11916-C-T, 14408-C-T, 18998-C-T, 23403-A-G, 25563-G-T, 29540-G-A) were from the COVID-19 epicenter in New York and neighboring states. Comprehensive phylogenetic analysis of the US-specific isolates, along with 2,000 randomly selected non-US isolates, revealed that the isolates fell exclusively in some major clades and were completely absent in the remaining clades (**Figure 1**). The mean number of isolates represented by each USA-specific haplotype was 11.4 ± 12.9 (range: 5–91) (**Supplementary Table S3**). Of note, 58 of the purely US-specific and the four nearly US-specific haplotypes consisted of 715 US isolates all had the globally dominant 23403-A-G (D614G) mutation (Korber et al., 2020). The 8782-C-T (orf1ab, synonymous) and 28144-T-C (orf8:p.Leu84Ser) variants were mutually exclusive with D614G, and co-occurred in 25 haplotypes that accounted for a total of 551 US isolates.

State-Specific Haplotypes

State-specific haplotypes were identified for 12 states based on the sequences of 613 out of 6,356 US isolates (9.6%). Further, seven US-specific haplotypes were almost exclusively found in isolates from a single state, where the few exclusivity-violating isolates were from neighboring states. Washington state had the most private haplotypes (24 haplotypes and 367 isolates), followed by California (9 haplotypes and 63 isolates) and Utah (5 haplotypes and 44 isolates). The number of haplotypes increased over time as new variants were continuously acquired, but the newly emerged haplotypes were confined within these states to accumulate such that the percentage of isolates that carry the USA- and state-private variants and haplotypes increase (**Supplementary Table S4**). Two of the California specific haplotypes are notable. The 9-isolate 491-G-A,14940-A-G haplotype group and its single-marker ancestral haplotype (14940-A-G with 5 isolates) were exclusively present in California between March 31, 2020 and May 1, 2020. 491-G-A is a missense variant, p.Ala76Thr, in the orf1ab gene. The 15-isolate 25692-C-T haplotype group is similarly interesting in the sense that these haplotypes are

[4]https://mummer4.github.io/

[5]https://www.gisaid.org/

[6]https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/

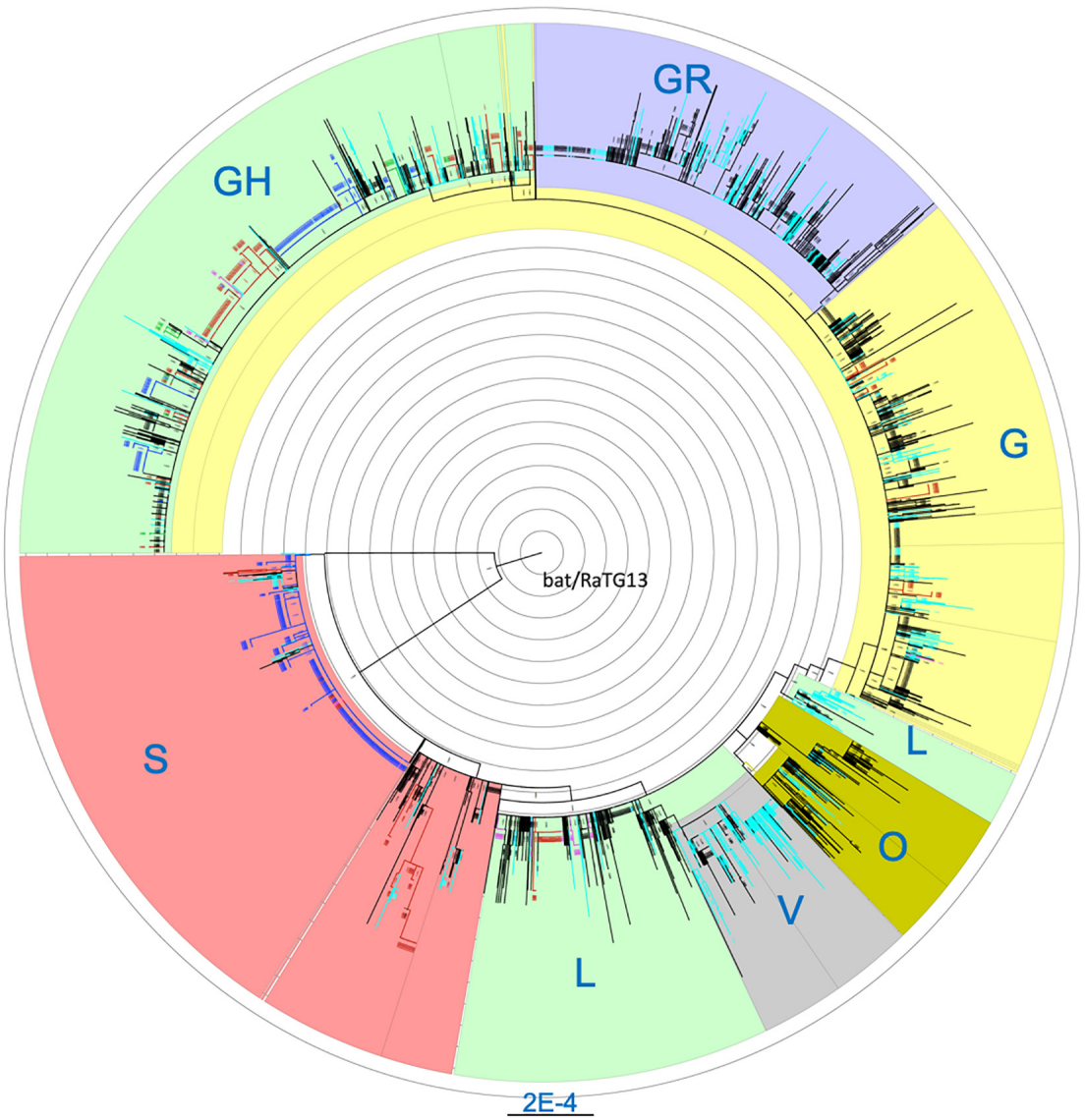


FIGURE 1 | Maximum likelihood phylogenetic tree of isolates carrying US-specific haplotypes and 2,000 randomly selected non-US isolates. Blue, isolates with Washington-specific haplotypes; Pink, Children's Hospital Los Angeles (CHLA) and other California isolates; Green, New York-specific, Red, other states' isolates with USA-specific haplotypes; Light blue, UK and Australia isolates; Black, Other non-US isolates. Blue labels, the clade names. A total of 1107 isolates from US-specific haplotypes were included in the phylogenetic analysis. Non-US isolates belonging to any haplotypes were randomly sampled at about 10% to reach 2000 isolates. Each branch in the phylogenetic tree may represent a group of isolates. The tree is rooted at the outgroup MN996532 (EPI_ISL_402131, bat/Yunnan/RaTG13/2013).

relatively "ancestral" with only 1-base difference from the reference isolate genome (NC_045512.2), but they were recently seen in late April, after the inception of the safer at home policy in California. This is in contrast with the dominant haplotypes in the US that were more distant descendants of NC_045512.2, with at least 3 and frequently more than 10 variants compared to the reference. This is suggestive of containment of early infections in California and limited spread to other states, likely again because of the early response to the pandemic from the state of California.

On the national level, one major haplotype (8782-C-T, 17747-C-T, 17858-A-G, 18060-C-T, 28144-T-C) had 317 member isolates, where 315 (99.4%) were from the US (311, 98.0%) and neighboring Canada locations (4, 1.3%). With the exception of two isolates from Australia, there were no isolates from outside North America. It is noted that this haplotype lacks the dominant D614G mutation prevalent in Europe. The first reported USA COVID-19 case in mid-January, haplotype (3 variants 8782-C-T, 18060-C-T, 28144-T-C), is the more remote ancestral haplotype. Three cases from Washington collected around January 18 shared this 3-variant haplotype. The isolates were continuously present from late February through April 2020, with predominance found in the western states, including

Washington and California, compared with isolates from the east coast. The potential immediate ancestral haplotypes inferred with CHLA CARD Genome Tracker were also from US and Canada isolates but they were sampled at later dates (Shen et al., 2020b). This provided further evidence of reduced state-to-state and coast-to-coast transmissions within the United States.

DISCUSSION

Based on the genomic analysis of all published SARS-CoV-2 sequences to date, safer-at-home measures have been very effective at reducing the spread of SARS-CoV-2, especially in preventing inter-state transmission of this highly infectious virus. The most state-private haplotypes were seen in the states of Washington and California where the earliest COVID-19 cases were reported and early safer-at-home orders were implemented. Persistent implementation of these measures would clearly lead to reduced spread of the COVID-19 pandemic over time. Indeed, in our recent re-analysis of SARS-CoV-2 genome data (August 3rd, 2020), we identified consistent, and even more prominent trend of localized haplotype patterns in California and other states. As an example, 3,048 of the 3,492 (87.3%) isolates from California carried state-specific haplotypes not seen in other states of USA.

On the other hand, a virus with a novel haplotype is not necessarily a new strain of the virus. The viral variant and haplotype analysis described here may prove to be critical, however, if a more transmissible and more deadly strain of SARS-CoV-2 emerges over time. Further studies will likely determine viral haplotypes, in the context of host factors, that may be associated with disease severity, response to treatment, or utility of vaccines for disease prevention.

AUTHOR'S NOTE

This manuscript has been released as a preprint at MedRxiv (Shen et al., 2020a).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://covid19.cpmbiodev.net/covid19/index.php.

AUTHOR CONTRIBUTIONS

All authors made major contributions to the conceptualization of the study, data interpretation, as well as the development of the manuscript.

ACKNOWLEDGMENTS

All data used in the study were extracted from the aforementioned public repositories, namely GISAID (https://www.gisaid.org/), NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), and the China National Center for Bioinformation (CNCB) 2019 nCoV Resource (https://bigd.big.ac.cn/ncov/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.573430/full#supplementary-material

REFERENCES

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contributions to global health. *Glob. Chall.* 1, 33–46. doi: 10.1002/gch2.1018

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* [Preprint], doi: 10.1101/2020.04.29.069054

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14:e1005944. doi: 10.1371/journal.pcbi.1005944

Shen, L., Dien Bard, J., Biegel, J. A., Judkins, A. R., and Gai, X. (2020a). Comprehensive genome analysis of 6,000 USA SARS-CoV-2 isolates reveals haplotype signatures and localized transmission patterns by state and by country. *medRxiv* [Preprint], doi: 10.1011/2020.05.23.20110452

Shen, L., Maglinte, D., Ostrow, D., Pandey, U., Bootwalla, M., Ryutov, A., et al. (2020b). Children's hospital los angeles COVID-19 analysis research database (CARD) - A resource for rapid SARS-CoV-2 genome identification using interactive online phylogenetic tools. *bioRxiv* [Preprint], doi: 10.1101/2020.05.11.089763

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

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3

Zhao, W. M., Song, S. H., Chen, M. L., Zou, D., Ma, L. N., Ma, Y. K., et al. (2020). The 2019 novel coronavirus resource. *Yi Chuan.* 42, 212–221. doi: 10.16288/j.yczz.20-030

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

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.

Copyright © 2020 Shen, Dien Bard, Biegel, Judkins and Gai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 September 2020
doi: 10.3389/fmicb.2020.582437

Role of BgaA as a Pneumococcal Virulence Factor Elucidated by Molecular Evolutionary Analysis

Masaya Yamaguchi[1], Moe Takemura[1,2], Kotaro Higashi[1], Kana Goto[1†], Yujiro Hirose[1], Tomoko Sumitomo[1], Masanobu Nakata[1†], Narikazu Uzawa[2] and Shigetada Kawabata[1]**

[1] Department of Oral and Molecular Microbiology, Osaka University Graduate School of Dentistry, Suita, Japan, [2] Department of Oral and Maxillofacial Surgery II, Osaka University Graduate School of Dentistry, Suita, Japan

Streptococcus pneumoniae is a major cause of pneumonia, sepsis, and meningitis. Previously, we identified a novel virulence factor by investigating evolutionary selective pressure exerted on pneumococcal choline-binding cell surface proteins. Herein, we focus on another pneumococcal cell surface protein. Cell wall-anchoring proteins containing the LPXTG motif are conserved in Gram-positive bacteria. Our evolutionary analysis showed that among the examined genes, *nanA* and *bgaA* had high proportions of codon that were under significant negative selection. Both *nanA* and *bgaA* encode a multi-functional glycosidase that aids nutrient acquisition in a glucose-poor environment, pneumococcal adherence to host cells, and evasion from host immunity. However, several studies have shown that the role of BgaA is limited in a mouse pneumonia model, and it remains unclear if BgaA affects pneumococcal pathogenesis in a mouse sepsis model. To evaluate the distribution and pathogenicity of *bgaA*, we performed phylogenetic analysis and intravenous infection assay. In both Bayesian and maximum likelihood phylogenetic trees, the genetic distances between pneumococcal *bgaA* was small, and the cluster of pneumococcal *bgaA* did not contain other bacterial orthologs except for a *Streptococcus gwangjuense* gene. Evolutionary analysis and BgaA structure indicated BgaA active site was not allowed to change. The mouse infection assay showed that the deletion of *bgaA* significantly reduced host mortality. These results indicated that both *nanA* and *bgaA* encode evolutionally conserved pneumococcal virulence factors and that molecular evolutionary analysis could be a useful alternative strategy for identification of virulence factors.

Keywords: ***Streptococcus pneumoniae*, molecular evolutionary analysis, pneumococcal cell wall-anchoring proteins, *bgaA*, *nanA***

OPEN ACCESS

Edited by:
Zhan Zhou,
Zhejiang University, China

Reviewed by:
Masamitsu Kono,
Wakayama Medical University, Japan
Anders P. Hakansson,
Lund University, Sweden

****Correspondence:***
Masaya Yamaguchi
yamaguchi@dent.osaka-u.ac.jp
Shigetada Kawabata
kawabata@dent.osaka-u.ac.jp

†Present address:
Kana Goto,
Department of Pediatric Dentistry, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan
Masanobu Nakata,
Department of Oral Microbiology, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima, Japan

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

Received: *12 July 2020*
Accepted: *01 September 2020*
Published: *24 September 2020*

Citation:
Yamaguchi M, Takemura M, Higashi K, Goto K, Hirose Y, Sumitomo T, Nakata M, Uzawa N and Kawabata S (2020) Role of BgaA as a Pneumococcal Virulence Factor Elucidated by Molecular Evolutionary Analysis. Front. Microbiol. 11:582437. doi: 10.3389/fmicb.2020.582437

INTRODUCTION

Streptococcus pneumoniae is one of most frequently isolated bacteria from community acquired pneumonia, sepsis, and bacterial meningitis (Ishiguro et al., 2013; CDC, 2019). The phylogenetic relationship of both 16S rRNA and a core set of 136 genes indicated that this pathogen belongs to mitis group of *Streptococcus* (Kawamura et al., 1995; Richards et al., 2014). *S. pneumoniae* is capable of importing various genes including antimicrobial resistance genes via horizontal transfer from

related species (Salvadori et al., 2019). In the United States of America, more than 30% of clinically isolated pneumococcal bacteria are resistant to one or more antibiotics (CDC, 2019). Currently, 23-valent pneumococcal polysaccharide and 13-valent pneumococcal-conjugated vaccines are in use in various countries. These vaccines prevent pneumococcal infections caused by vaccine-targeted serotype strains and inhibit the spread of drug resistant strains (CDC, 2019). On the other hand, the selective pressure imposed by vaccination has increased the emergence of non-vaccine serotype strains (Golubchik et al., 2012).

Recently, we applied a combination of evolutionary analysis and laboratory-based approaches to evaluate the functional significance of putative virulence factors (Yamaguchi et al., 2016, 2019a; Yamaguchi, 2018). As mutations in non-essential but important genes promote the selection of bacterial lineages in single species, genes that undergo considerable negative selection would be important for the survival and/or success of the species in its host and/or the environment. Thus, a molecular evolutionary approach enables us to estimate the contributions of bacterial proteins to species success throughout its life cycle. We have previously focused on the evolutionary selective pressures on choline-binding proteins (CBPs) in *S. pneumoniae* (Yamaguchi et al., 2019a). CBPs are localized to the cell surface by binding to phosphoryl choline on the cell wall. As cell surface proteins are easily and directly accessible to the external environment, these proteins could represent attractive antigen candidates for vaccine development. Our analysis revealed that CbpJ contributes to evasion of host neutrophil-mediated killing in pneumococcal pneumonia. This is surprising as CbpJ has no known functional domains apart from signal sequences and choline-binding repeats. While there are also other types of pneumococcal cell surface proteins, their degree of evolutionary conservation remains unknown.

For this study, we focused on different motif involved in pneumococcal cell surface localization, LPXTG, which is associated with cell wall-anchoring (Lofling et al., 2011). LPXTG-containing proteins are covalently attached to the cell wall, and at least some of these proteins have been identified as multi-functional proteins (Yamaguchi et al., 2008; Uchiyama et al., 2009; Dalia et al., 2010; Lofling et al., 2011). Our evolutionary analysis indicated that the *nanA* and *bgaA* genes are under considerable negative selection pressure. The *nanA* gene encodes pneumococcal cell surface-localized exo-α-sialidase (NanA) that hydrolyzes α2-3-, α2-6-, and α2-8-linkages of *N*-acetylneuraminic acid residues, and *bgaA* encodes exo-β-galactosidase (BgaA) that hydrolyzes β1-4-linkages of galactose residues, respectively (Hobbs et al., 2018). These glycosidases contribute to biofilm formation in glucose-poor but galactose-rich environments such as the mouse nasopharynx (Blanchette et al., 2016). The glycosidases also disrupt complement deposition and reduce opsonophagocytic killing through catalytic activities (Dalia et al., 2010). In mouse intravenous infection, NanA contributes to pneumococcal invasion into the host central nervous system by aiding penetration through the blood–brain barrier (Uchiyama et al., 2009). However, whether BgaA functions as a virulence factor *in vivo* remains unknown. Thus, we performed a phylogenetic analysis and mouse intravenous infection assay to address this question.

MATERIALS AND METHODS

Phylogenetic and Evolutionary Analyses

The tBLASTn search was used to identify homologs and orthologs of genes that encode cell wall-anchoring proteins (Gertz et al., 2006). Phylogenetic and evolutionary analyses were performed as previously described, with minor modifications (Yamaguchi et al., 2016, 2017, 2019a). Briefly, the sequences were aligned by codon using Phylogear2 (Tanabe, 2008), MAFFT v.7.221 with an L-INS-i strategy (Katoh and Standley, 2013), and Jalview (Waterhouse et al., 2009). Conserved common codons were used for further phylogenetic analysis. The best-fitting codon evolutionary models for MrBayes and RAxML analyses were determined using Kakusan4 (Tanabe, 2011). Bayesian Markov chain Monte Carlo analyses were performed using MrBayes v.3.2.5 or v.3.2.6 (Ronquist et al., 2012), and 2–8 $\times$ 10^6 generations were sampled. To validate phylogenetic inferences, maximum likelihood phylogenetic trees with bootstrap values were generated with RAxML v.8.1.20 (Stamatakis, 2014). Phylogenetic trees were visualized using FigTree v.1.4.4 (Rambaut, 2018). Evolutionary analyses were performed based on aligned orthologous regions of genes that encode cell wall-anchoring proteins and Bayesian phylogenetic trees with a two-rate fixed-effects likelihood function in the HyPhy software package (Pond et al., 2005). For the evolutionary analyses, the level of statistical significance was set at $P < 0.1$ with the HyPhy default setting.

TIGR4 BgaA protein structure was visualized using PyMOL 2.4[1]. The PDB ID is 4CU6 (Singh et al., 2014). The domain structures were identified using MOTIF and Pfam (Kanehisa and Goto, 2000; Finn et al., 2014).

Bacterial Strains and Construction of Mutant Strains

S. pneumoniae strains were cultured at 37°C in Todd-Hewitt broth (BD Biosciences, Franklin Lakes, NJ, United States) supplemented with 0.2% yeast extract (THY; BD Biosciences). Spectinomycin (Wako Pure Chemical Industries, Osaka, Japan) was added to the medium to a concentration of 120 μg/mL for mutant selection and maintenance. The *S. pneumoniae* TIGR4 isogenic *bgaA* (Δ*bgaA*) mutant strain was generated as previously described (Mori et al., 2012; Yamaguchi et al., 2019b). Briefly, the upstream region of *bgaA*, an *aad9* cassette, and the downstream region of *bgaA* were combined by PCR using the primers summarized in **Supplementary Table 1**. The PCR product was transformed with synthesized CSP2 to construct the mutant strains by double-crossover recombination (Bricker and Camilli, 1999). The mutation was confirmed by site-specific PCR with isolated genomic DNA from the mutant strains. For growth measurement, overnight cultures of each strain were back-diluted 3:100 into fresh THY and grown at 37°C. Growth

[1]https://pymol.org/2/

was monitored by measuring the OD_{600} values every 30 min. The starting point was set at an OD_{600} value of approximately 0.1. The experiment was repeated three times and the data is provided as **Supplementary Data 1**.

Mouse Intravenous Infection Assays

All mouse experiments were conducted in accordance with animal protocols approved by the Animal Care and Use Committee of Osaka University Graduate School of Dentistry (28-002-0). The mouse infection assay was performed as previously described (Hirose et al., 2018; Yamaguchi et al., 2019a). Briefly, CD-1 mice (Slc:ICR, 6 weeks, female) were infected by tail vein injection with 1×10^6 CFUs of *S. pneumoniae*. Mouse survival was checked twice daily for 14 days. The experiment was repeated three times and the data is provided as **Supplementary Data 2**. The pooled data for the three experiments was compared using a log-rank test. Statistical analysis was performed using Prism v.7.0d or v.8.4.2 software (GraphPad, Inc., La Jolla, CA, United States). The level of significance for differences between groups was set at $P < 0.05$.

RESULTS

Evolutionary Selective Pressures on Genes Encoding Cell Wall-Anchoring Proteins

The tBLASTn function was used to search pneumococcal genomes for genes encoding cell wall-anchoring proteins (**Figure 1** and **Supplementary Table 2**). While no genes were found to be conserved as intact open reading frames in all strains, all genes, save for *pclA* and *psrP,* were present in most strains. In particular, although the *nanA* gene in the TIGR4 strain contains a frameshift mutation in C-terminal region, TIGR4 NanA shows sialidase activity, and is assumed to be secreted into milieu instead of being anchored to the cell wall (Gut et al., 2011).

To evaluate the degree of evolutionary conservation in cell wall-anchoring proteins, we performed molecular evolutionary calculations based on each phylogenetic relationship and the DNA sequences aligned by codon. The calculated selective pressure for each gene is summarized in **Table 1**. The percentage of codons that are negatively selected for is visualized in **Figure 1B**. There was negative selection for over 11% of total codons in *nanA* and *bgaA*. This contrasts with less than 5% of total codons in most other genes, indicating that these two genes play an important role in the success of *S. pneumoniae* species. This same tendency was observed in our previous analysis on pneumococcal CBPs using the same genome sequences. Specifically, in the previous study, more than 13% of codons in the top two genes, *cbpJ* and *lytA*, were negatively selected (Yamaguchi et al., 2019a). Further, we previously reported that the *pfbA* gene showed high specificity to *S. pneumoniae* species and had a low level of sequence diversity (Yamaguchi et al., 2019b). Interestingly, our evolutionary analysis also indicated that the *pfbA* gene is under relaxed selective pressure.

Phylogenetic Relationships of the *bgaA* Gene

Evolutionary analysis indicated that the top two genes, *nanA* and *bgaA,* had high percentage of codons that were under negative selection pressure. We have previously reported that streptococcal *nanA* orthologs diverged into two major groups, with one group consisting of *Streptococcus mitis, Streptococcus intermedius* and *S. pneumoniae*, and the other consisting of *Streptococcus agalactiae* and the *Streptococcus iniae* groups (Yamaguchi et al., 2016). However, the phylogenetic relationships of *bgaA* have not been previously described. We used tBLASTn to search for *bgaA*-homologs, and performed Bayesian and maximum likelihood phylogenetic analyses (**Figure 2** and **Supplementary Figure 1**). The tBLASTn search of the NCBI Nucleotide collection database showed that the *bgaA* gene homologs were identified in various streptococcal species and other Gram-positive bacteria of the phylum Firmicutes, including genus *Clostridium* and *Bacillus.* The β-galactosidase genes of these Gram-positive bacteria were used to root the phylogenetic trees. Bayesian and maximum likelihood phylogenetic analyses produced almost identical trees. The *bgaA* genes in *S. pneumoniae* and *S. pseudopneumoniae* had small genetic distances and formed an independent cluster within a cluster of streptococcal strains. In contrast, orthologous genes in other streptococcal species were genetically diverse. One of the *bgaA* orthologs from the *Streptococcus gwangjuense* strain ChDC B345 belonged to the pneumococcal clusters. As this gene is distinct from other *S. gwangjuense* genes, there is a possibility that it had been obtained from *S. pneumoniae via* horizontal gene transfer.

Evolutionarily Conserved Catalytic Residues Contribute to the Conformation of BgaA Active Site

The domain structures and amino acid residues of TIGR4 BgaA that are under negative selection are shown in **Figures 3A–D** and **Supplementary Figure 2**. BLAST search showed that the TIGR4 BgaA amino acid sequence did not have high similarity with human proteins (**Supplementary Table 3**). BgaA contains glycosyl hydrolase domains in its N-terminus, and most evolutionarily conserved residues were present, from the glycosyl hydrolase sugar binding domain to the first bacterial Ig-like domain (**Figures 3A,B** and **Supplementary Figure 2**). All evolutionarily conserved proline located on loop regions (**Figure 3C**). Since proline has restricted phi-psi space that arise from the 5-membered ring and stabilize protein structure (Morris et al., 1992), the residues may contribute to the conservation of BgaA catalytic region via stabilization of those loop structures. The active site of BgaA contains 15 catalytic residues, R288, H450, H484, E564, D599, R602, F603, Y624, W685, W708, Y713, E716, T718, H721, and F733 (Singh et al., 2014). All catalytic residues were present in comparable residues encoded by commonly conserved codons in pneumococcal species. Interestingly, eight catalytic residues shown in **Figure 3D** were evolutionarily conserved ones. The other seven catalytic residues were flanked (R288, H450, W685, and W708) by or located (E564, D599, and Y713) within two residues from evolutionarily conserved

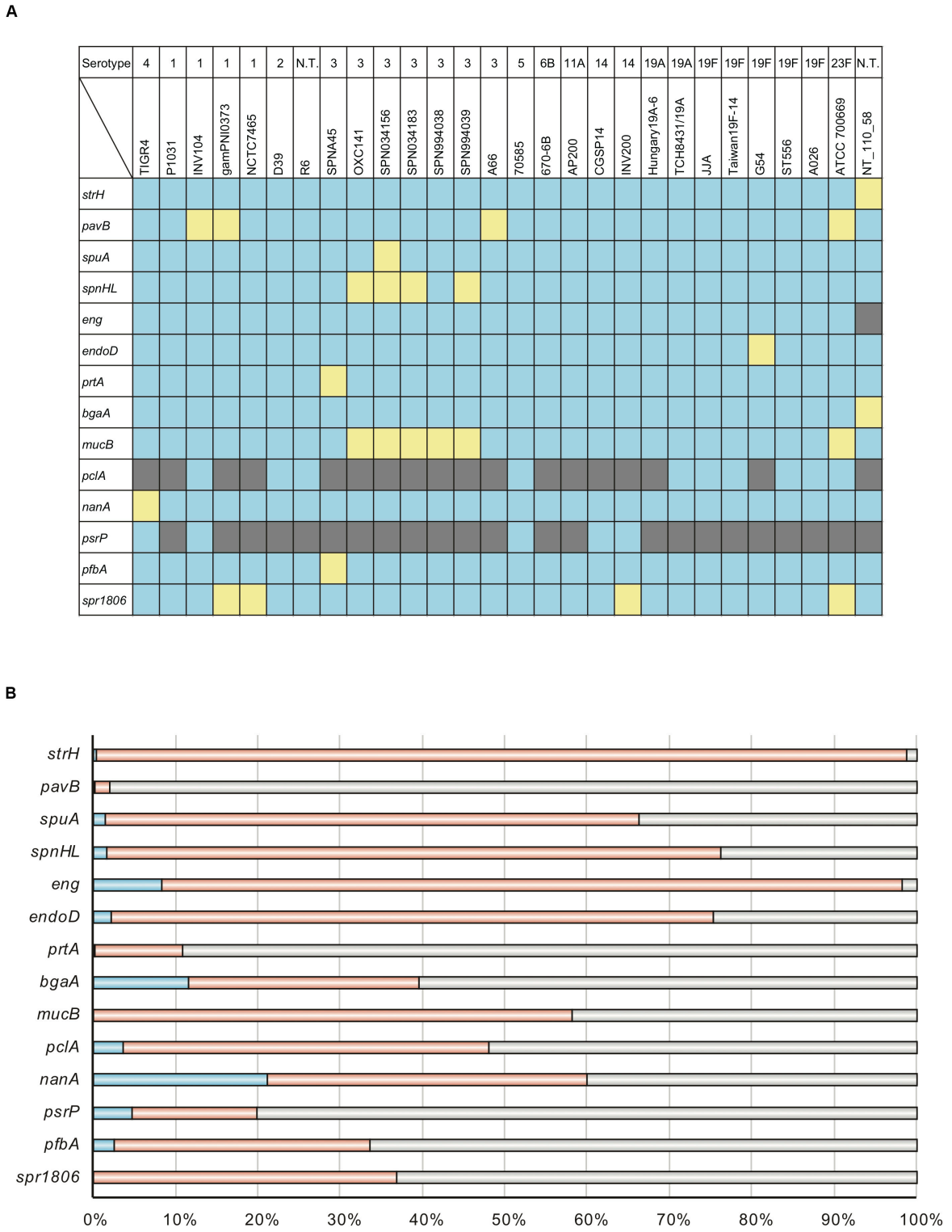


FIGURE 1 | Distribution of genes that encode cell wall-anchoring proteins and percentage of codons that are evolutionarily conserved for these genes. **(A)** Distribution of genes that encode cell wall-anchoring proteins from pneumococcal strains. The gene locus tag numbers are summarized in **Supplementary Table 2**. Blue, yellow, and gray represent the presence, pseudogenization, and absence of genes, respectively. **(B)** Codons of genes encoding cell wall-anchoring proteins evolving under purifying selection were identified using HyPhy software with phylogenetic trees and aligned sequences. Blue, orange, and gray represent the percentage of codons under purifying selection, comparable common codons, and incomparable codons, respectively. The actual numbers and other parameters are listed in **Table 1**.

TABLE 1 | Evolutionary analyses of genes that encode cell wall-anchoring proteins.

Genes	Number of sequences[1]	dN/dS	Coverage of comparable codons relative to whole protein in TIGR4	Codons evolving under positive selection	Codons evolving under purifying selection	% of codons under purifying selection relative to total codons
strH	23	0.561	98.781% (1297/1313)	0.154% (2/1297)	0.463% (6/1297)	0.457%
pavB	25	0.181	1.981% (17/858)	0% (0/17)	11.765% (2/17)	0.233%
spuA	22	0.151	66.354% (850/1281)	0% (0/850)	2.353% (20/850)	1.561%
spnHL	24	0.343	76.289% (814/1067)	0.369% (3/814)	2.211% (18/814)	1.687%
Eng	20	0.145	98.303% (1738/1768)	0.230% (4/1738)	8.458% (147/1738)	8.314%
endoD	23	0.243	75.361% (1251/1660)	0.080% (1/1251)	2.878% (36/1251)	2.169%
prtA	21	0.013	10.836% (232/2141)	0% (0/232)	3.017% (7/232)	0.327%
bgaA	22	0.185	39.526% (883/2234)	0.113% (1/883)	29.332% (259/883)	11.594%
mucB	16	1.147	58.128% (118/203)	0% (0/118)	0% (0/118)	0.000%
pclA	6	0.328	48.002% (1225/2552)[2]	0.082% (1/1225)	7.592% (93/1225)	3.644%
nanA	23	0.170	60.020% (599/998)[2]	0.334% (2/599)	35.392% (212/599)	21.242%
psrP	5	0.203	19.845% (948/4777)	0.633% (6/948)	24.367% (231/948)	4.836%
pfbA	19	0.423	33.568% (238/709)	0% (0/238)	7.983% (19/238)	2.680%
spr1806	19	5.172	36.937% (82/222)	0% (0/82)	0% (0/82)	0.000%

[1]Sequences with 100% identity were considered to be the same sequence; [2]compared to D39. Evolutionary analysis was performed by Bayesian inference of aligned sequences from complete genomes of S. pneumoniae using the two-rate fixed-effects likelihood function in the HyPhy software package. dN/dS is the ratio of non-synonymous to synonymous changes overall in analyzed genes. Individual codons with a statistically significant signature were also tallied and are expressed as a percentage of the total number of codons included in the analysis.

residues. These results indicated that evolutionarily conserved residues contribute to the conformation of the BgaA active site.

BgaA Deficiency Decreases Pneumococcal Pathogenicity in a Mouse Sepsis Model

NanA has been identified as a multi-functional virulence factor. It contributes to pneumococcal biofilm formation, adhesion to and invasion of host epithelial and endothelial cells, inducing excessive host inflammatory responses, and resistance to opsonophagocytosis (King et al., 2006; Uchiyama et al., 2009; Dalia et al., 2010; Chang et al., 2012; Blanchette et al., 2016). BgaA also has been reported to contribute to pneumococcal adherence to host epithelial cells, host immune evasion, and *in vivo* biofilm formation (Dalia et al., 2010; Limoli et al., 2011; Singh et al., 2014; Blanchette et al., 2016). However, several studies reported that deletion of *bgaA* had limited effects on *S. pneumoniae* in mouse colonization models (King et al., 2006; Limoli et al., 2011; Blanchette et al., 2016; Hobbs et al., 2018). It also remains unclear if the deletion of *bgaA* would significantly affect host survival rate in a mouse model of sepsis. Thus, we constructed a TIGR4 *bgaA* mutant (Δ*bgaA*) strain, and performed a mouse intravenous infection assay to compare host survival for TIGR4 wild type and Δ*bgaA* strains. These strains showed similar growth rates in THY medium (**Supplementary Figure 3** and **Supplementary Data 1**). In this infection model, we found that Δ*bgaA*-infected mice had a significantly higher survival rate compared to mice infected with the TIGR4 wild type strain (**Figure 4** and **Supplementary Data 2**). This result indicates that BgaA functions as a virulence factor in a mouse sepsis model.

DISCUSSION

In this study, we investigated the percentage of codons in genes that encode cell wall-anchoring proteins that are under negative selection in *S. pneumoniae* species. Over 11% of codons in the *nanA* and *bgaA* genes had significant negative selection. As NanA is a well-studied virulence factor in *S. pneumoniae*, we chose to focus on the *bgaA* gene that encodes β-galactosidase. Phylogenetic analysis indicated that the *bgaA* genes in *S. pneumoniae* and *S. pseudopneumoniae* had small genetic distances and formed a distinct cluster within streptococcal *bgaA* orthologs. In addition, eight catalytic residues were evolutionarily conserved, and the other seven catalytic residues were located near evolutionarily conserved ones. We also demonstrated using a mouse intravenous infection model that BgaA contributes to pneumococcal pathogenesis *in vivo*. In combination with our previous work on pneumococcal CBPs, these results suggest that the degree of evolutionary conservation could be an effective parameter for estimating the importance of bacterial cell surface proteins (Yamaguchi et al., 2019a).

While a classical molecular microbiology approach involves analysis of a few representative strains, an evolutionary analysis approach is able to reflect importance in a species based on analysis of a few dozen to 1000s of genome sequences. For genes that encode pneumococcal CBPs, we have recently shown that more than 13% of codons in *cbpJ* and *lytA* genes are negatively selected, and CbpJ acts as a novel virulence factor in pneumococcal pneumonia both *in vitro* and *in vivo* (Yamaguchi et al., 2019a). In this study, we observed negative selective pressure for over 11% of codons in *nanA* and *bgaA*. In contrast, this figure was 3% of codons for *pfbA*. We have previously reported that PfbA is pneumococcal cell surface protein that

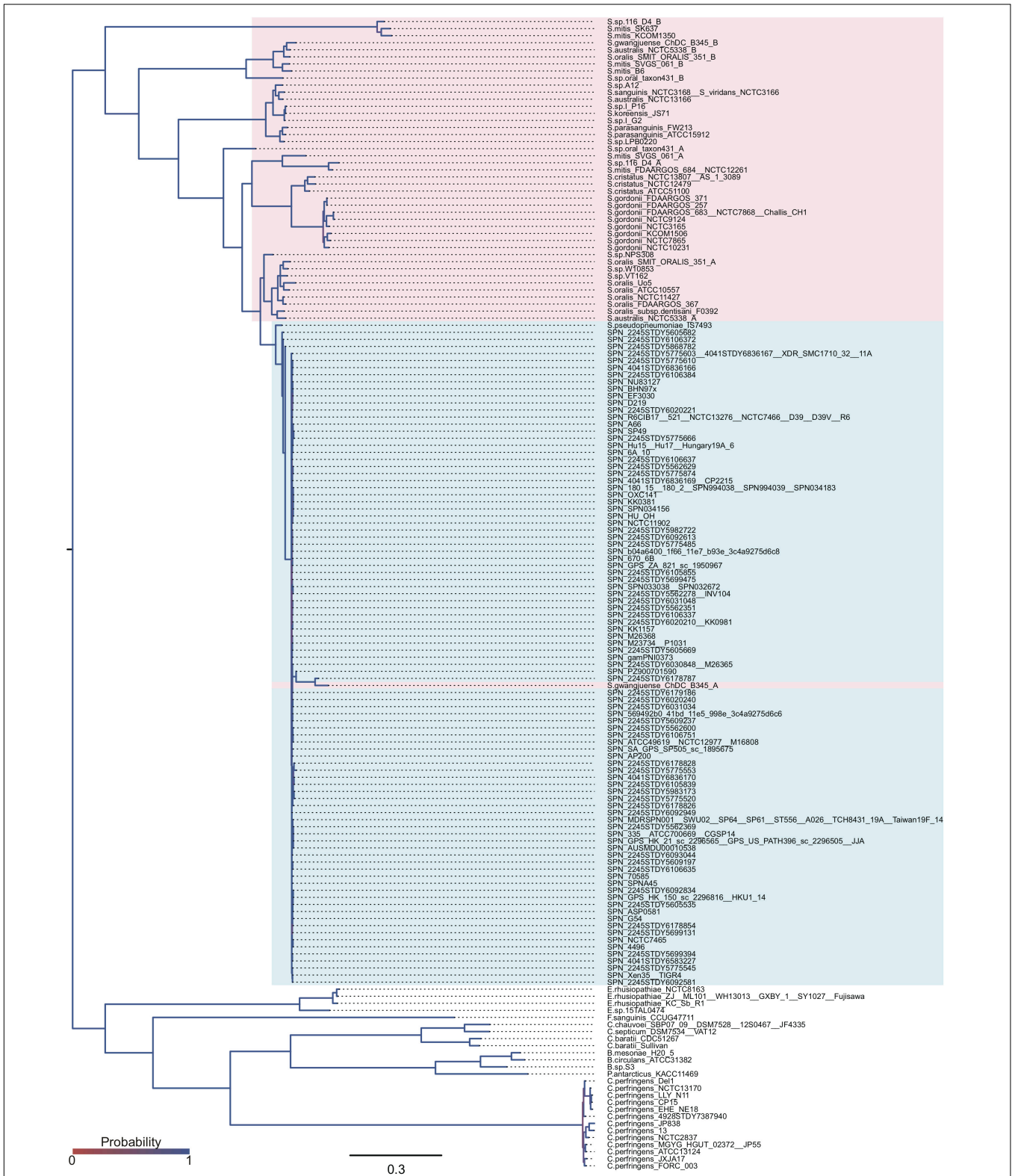


FIGURE 2 | Bayesian phylogenetic analysis of the *bgaA* gene. The codon-based Bayesian phylogenetic relationship was calculated using the MrBayes program. Strains with identical sequences are listed on the same branch. *S. pneumoniae* and *S. pseudopneumoniae bgaA* genes are shaded in cyan. Other streptococcal *bgaA* ortholog genes are shaded in magenta. The color gradation of phylogenetic tree represents posterior probability. The scale bar indicates nucleotide substitutions per site.

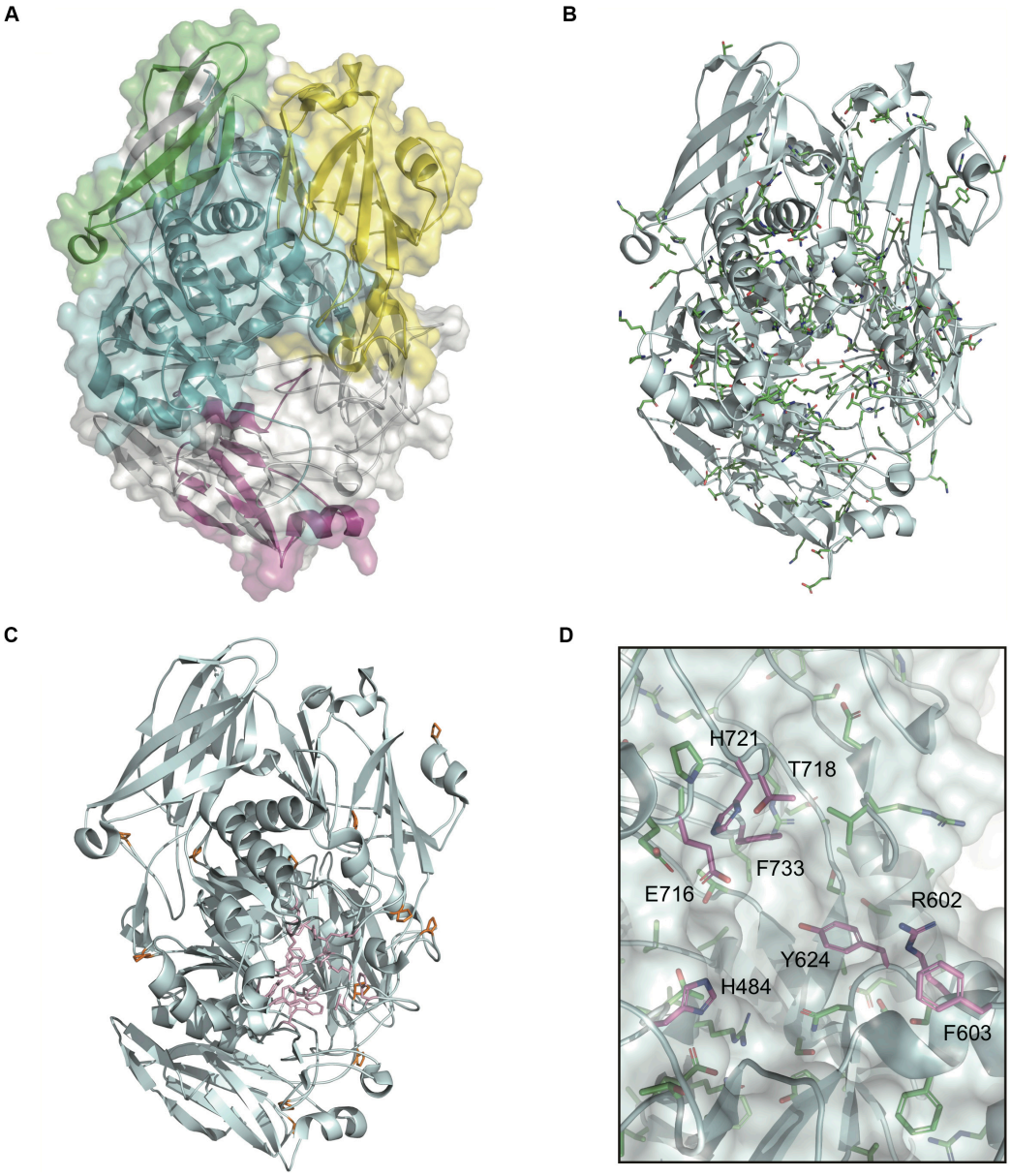


FIGURE 3 | Evolutionarily conserved residues in the BgaA catalytic region. **(A)** Pictorial representation of the structure of TIGR4 BgaA catalytic region (PDB ID: 4CU6). Glycosyl hydrolase family 2 sugar binding domain, glycosyl hydrolase family 2 domain, glycosyl hydrolase family 2 TIM barrel domain, and DUF4982 are colored in yellow, green, cyan, and magenta, respectively. **(B)** The side chains of evolutionarily conserved residues are shown as colored stick models. Carbon, nitrogen, and oxygen are shown as green, blue, and red, respectively. **(C)** BgaA active site is shown as pink stick models, and evolutionary conserved proline residues are in orange. **(D)** The active site of BgaA. The carbon of the evolutionarily conserved catalytic residues is shown in purple (H484, R602, F603, Y624, E716, T718, H721, and F733). The carbon of the evolutionarily conserved non-catalytic residues is shown in green. Nitrogen and oxygen are shown as blue and red, respectively.

forms a right-handed parallel β-helix and interacts with human fibronectin, plasmin, plasminogen, albumin, hemoglobin, and fibrinogen (Yamaguchi et al., 2008; Beulin et al., 2014, 2017; Radhakrishnan et al., 2018; Sankar et al., 2020). A BLAST search and phylogenetic analysis showed that *pfbA* is highly conserved in *S. pneumoniae* among mitis group *Streptococcus* (Yamaguchi et al., 2019b). In addition, *in vitro* assays revealed that PfbA functions as an adhesin and invasin for human epithelial cells, is a TLR2 ligand, and an anti-phagocytic factor for human neutrophils (Yamaguchi et al., 2008, 2019b). On the other hand, the deletion of *pfbA* in *S. pneumoniae* had no significant effect on host mortality in a mouse pneumonia model, and enhanced pneumococcal pathogenesis in a mouse model of sepsis (Yamaguchi et al., 2019b). With regards to pneumococcal cell surface proteins, its degree of evolutionary conservation shows good correlation with results in a mouse infection

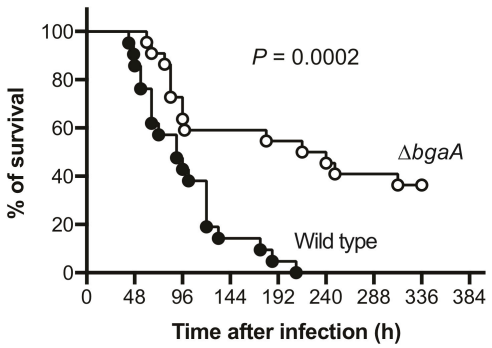


FIGURE 4 | Deficiency of *bgaA* decreases pneumococcal virulence in a mouse model of sepsis. Mice were infected by intravenous injection with 1×10^6 colony forming units (CFU) of *S. pneumoniae* TIGR4 wild type ($n = 21$), or Δ*bgaA* ($n = 22$), and survival was monitored for 14 days. The difference between infected mouse groups was compared using a log-rank test. The data were obtained from three independent experiments.

model. Although an evolutionary analysis approach has some limitations, including the fact that it is unable to identify diverse virulence factors within a species (Yamaguchi et al., 2019a), this approach could also be an effective alternative strategy for identification of common virulence factors. Furthermore, in this study, evolutionary analysis with protein structure information open the possibility that the residues under negative selection are important for the protein function and/or structure. Increasing availability and accessibility of bacterial genomic information would allow us to combine evolutionary analysis and laboratory-based approaches to study various bacterial species and proteins.

BgaA is regarded as a multi-functional protein and putative virulence factor; however, BgaA plays a limited or inconsequential role in *in vivo* colonization (King et al., 2006; Burnaugh et al., 2008; Dalia et al., 2010; Limoli et al., 2011; Singh et al., 2014; Blanchette et al., 2016; Hobbs et al., 2018). Here, we revealed that *bgaA* deficiency significantly reduced mortality in a mouse model of blood infection. As the host bloodstream is a glucose-rich environment, the ability to utilize host galactose as an alternative carbon source would not be crucial to pneumococcal survival (Blanchette et al., 2016). BgaA inhibits complement deposition, and consequently opsonophagocytosis, by cleaving *N*-glycans on host glycoproteins that are involved in the complement cascade (Dalia et al., 2010). As such, our observation that the Δ*bgaA* strain reduced host mortality in a mouse blood infection model may be explained by an inability to evade host complement deposition and phagocytosis. Although NanA, BgaA, and StrH have been reported to inhibit C3b deposition to the same extent (Dalia et al., 2010), the evolutionary approach indicated that high rates of *nanA* and *bgaA* codons are under negative selection, while *strH* is not. This discrepancy may arise from specificities of the glycosidases. For example, *N*-linked glycans are cleaved by NanA, BgaA, and StrH sequentially. NanA cleaves terminal sialic acids, after which BgaA cleaves terminal, or NanA-exposed galactose, since sialic acids are commonly linked to the C-3 or C-6 positions of galactose. Subsequently, StrH reportedly cleaves host glycans after NanA and BgaA cleavage (Hobbs et al., 2018). Therefore, one possible hypothesis is that the cleavage order of these glycosidases affects their relative importance. Specifically, StrH may function as a complementary glycosidase since NanA and BgaA may provide sufficient sialic acid, and galactose as alternative carbon sources in glucose-poor environments. Alternatively, the additional roles of NanA and BgaA may simply be more important for pneumococcal survival in the host. However, further studies are required to elucidate the precise role of BgaA in sepsis.

Although, in the current study, we focused on the top two evolutionary conserved genes, the third gene, *eng*, may also serve as an important virulence factor since 8% of *eng* codons are negatively selected. The *eng* gene encodes Eng, endo-α-*N*-acetylgalactosamidase, which specifically cleaves core-1 *O*-linked glycans (Brooks and Savage, 1997; Marion et al., 2009). Although Eng reportedly contributes to the colonization of mouse airway, its specific role in pneumococcal pathogenesis, and whether it affects host mortality, remain unclear. Hence, *eng* is also an attractive target for further investigations.

Emerging antimicrobial resistance of *S. pneumoniae* and serotype replacement after the introduction of pneumococcal vaccines are serious global challenges (O'Neill, 2016; CDC, 2019). A potential solution would involve developing a novel vaccine with a common antigen. *S. pneumoniae* has the ability to import genes from related species and undergo recombination (Kilian et al., 2014; Kilian and Tettelin, 2019). Thus, to minimize the possibility of selective pressure-mediated antigenic variation, a multi-valent vaccine would be superior to a monovalent vaccine. Our evolutionary analysis showed that four pneumococcal cell surface proteins are evolutionarily conserved. A combination of intact or truncated LytA, CbpJ, NanA, and BgaA may be attractive antigen candidates for the development of a universal pneumococcal vaccine. Several groups have already reported that individually, NanA and LytA work as protective antigens in mouse infection models (Berry et al., 1989; Lock et al., 1992; Long et al., 2004; Tong et al., 2005; Yuan et al., 2011; Janapatla et al., 2018). On the other hand, as non-pathogenic native microflora and other *S. mitis* group species also contain some of these genes, it is necessary to examine the possibility that the vaccine may alter the composition of oral and/or lung microbiomes. Evaluation of potential side effects on host microbiome, alongside characterization of the immunogenicity of antigens and efficacy, would help to guide the design of a novel vaccine.

PspA is a promising vaccine candidate for pneumococcal infection, as this vaccine is generally assumed to be multi-valent since PspA is a highly variable protein under positive selection (Piao et al., 2014; Chen et al., 2015; Cornick et al., 2017; Yamaguchi et al., 2019a). Hence, host immune systems, including the humoral immune response, would select for pneumococcal PspA. This may prove advantageous for a vaccine candidate since the presence of positive selection indicates that the protein is capable of inducing host acquired immunity. At the same time, the variety of PspA reflects that *S. pneumoniae* evolves to evade host acquired immunity through obtaining PspA mutations. Thus, the introduction of PspA vaccine would induce novel selective pressure causing "PspA replacement," referring to an increase in vaccine-uncovered type PspA, as well

as serotype replacement by pneumococcal polysaccharide capsule vaccines. It would, therefore, be of importance to determine whether the multi-valent PspA vaccine effectively covers the selective pressure-extended PspA variety. Indeed, two different selective pressures for polysaccharide and PspA vaccines may overcome the replacement issue. Hence, during pandemics of emerging infectious diseases, such as COVID-19, an effective strategy may be to develop novel vaccines that simultaneously target conserved proteins under positive and negative selections. However, proteins under positive selection are suitable as initial targets for rapid vaccine development as they have a strong probability to elicit antibodies in humans. Meanwhile, to address the evolution of pathogens, evolutionarily conserved proteins would serve as effective targets in the development of later vaccine candidates.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care and Use Committee of Osaka University Graduate School of Dentistry (28-002-0).

AUTHOR CONTRIBUTIONS

MY and SK designed the study. MY and KH performed bioinformatics analyses. MT performed the animal experiments. MY, KG, YH, TS, MN, and SK contributed to the experimental setup. MY wrote the manuscript. MT, KH, KG, YH, TS, MN, NU, and SK contributed to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported in part by the Japan Society for the Promotion of Science KAKENHI (grant numbers 16K15787, 17H05103, 18K19643, 19H03825, and 19K22710); SECOM Science and Technology Foundation; AMED (20wm0325001h0001); Takeda Science Foundation; MSD Life Science Foundation, Public Interest Incorporated Foundation; GSK Japan Research Grant; Asahi Glass Foundation; Kurata Memorial Hitachi Science and Technology Foundation; The Naito Foundation, and Kobayashi International Scholarship Foundation; and GlaxoSmithKline Japan. The funders had no role in the study design, data collection or analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Hikaru Kodama for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.582437/full#supplementary-material

REFERENCES

Berry, A. M., Lock, R. A., Hansman, D., and Paton, J. C. (1989). Contribution of autolysin to virulence of *Streptococcus pneumoniae*. *Infect. Immun.* 57, 2324–2330.

Beulin, D. S., Yamaguchi, M., Kawabata, S., and Ponnuraj, K. (2014). Crystal structure of PfbA, a surface adhesin of *Streptococcus pneumoniae*, provides hints into its interaction with fibronectin. *Int. J. Biol. Macromol.* 64, 168–173. doi: 10.1016/j.ijbiomac.2013.11.035

Beulin, D. S. J., Radhakrishnan, D., Suresh, S. C., Sadasivan, C., Yamaguchi, M., Kawabata, S., et al. (2017). *Streptococcus pneumoniae* surface protein PfbA is a versatile multidomain and multiligand-binding adhesin employing different binding mechanisms. *FEBS J.* 284, 3404–3421. doi: 10.1111/febs.14200

Blanchette, K. A., Shenoy, A. T., Milner, J. II, Gilley, R. P., McClure, E., Hinojosa, C. A., et al. (2016). Neuraminidase A-exposed galactose promotes *Streptococcus pneumoniae* biofilm formation during colonization. *Infect. Immun.* 84, 2922–2932. doi: 10.1128/IAI.00277-16

Bricker, A. L., and Camilli, A. (1999). Transformation of a type 4 encapsulated strain of *Streptococcus pneumoniae*. *FEMS Microbiol. Lett.* 172, 131–135.

Brooks, M. M., and Savage, A. V. (1997). The substrate specificity of the enzyme endo-alpha-*N*-acetyl-D-galactosaminidase from *Diplococcus pneumonia*. *Glycoconj. J.* 14, 183–190. doi: 10.1023/a:1018585604073

Burnaugh, A. M., Frantz, L. J., and King, S. J. (2008). Growth of *Streptococcus pneumoniae* on human glycoconjugates is dependent upon the sequential activity of bacterial exoglycosidases. *J. Bacteriol.* 190, 221–230. doi: 10.1128/JB.01251-07

CDC (2019). *Antibiotic Resistance Threats in the United States, 2019*. Atlanta, GA: CDC.

Chang, Y. C., Uchiyama, S., Varki, A., and Nizet, V. (2012). Leukocyte inflammatory responses provoked by pneumococcal sialidase. *mBio* 3:e00220-11. doi: 10.1128/mBio.00220-11

Chen, A., Mann, B., Gao, G., Heath, R., King, J., Maissoneuve, J., et al. (2015). Multivalent pneumococcal protein vaccines comprising pneumolysoid with epitopes/Fragments of CbpA and/or PspA elicit strong and broad protection. *Clin. Vaccine Immunol.* 22, 1079–1089. doi: 10.1128/CVI.00293-15

Cornick, J. E., Tastan Bishop, O., Yalcin, F., Kiran, A. M., Kumwenda, B., Chaguza, C., et al. (2017). The global distribution and diversity of protein vaccine candidate antigens in the highly virulent *Streptococcus pnuemoniae* serotype 1. *Vaccine* 35, 972–980. doi: 10.1016/j.vaccine.2016.12.037

Dalia, A. B., Standish, A. J., and Weiser, J. N. (2010). Three surface exoglycosidases from *Streptococcus pneumoniae*, NanA, BgaA, and StrH, promote resistance to opsonophagocytic killing by human neutrophils. *Infect. Immun.* 78, 2108–2116. doi: 10.1128/IAI.01125-09

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A., and Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* 4:41. doi: 10.1186/1741-7007-4-41

Golubchik, T., Brueggemann, A. B., Street, T., Gertz, R. E. Jr., Spencer, C. C., Ho, T., et al. (2012). Pneumococcal genome sequencing tracks a vaccine escape

variant formed through a multi-fragment recombination event. *Nat. Genet.* 44, 352–355. doi: 10.1038/ng.1072

Gut, H., Xu, G., Taylor, G. L., and Walsh, M. A. (2011). Structural basis for *Streptococcus pneumoniae* NanA inhibition by influenza antivirals zanamivir and oseltamivir carboxylate. *J. Mol. Biol.* 409, 496–503. doi: 10.1016/j.jmb.2011.04.016

Hirose, Y., Yamaguchi, M., Goto, K., Sumitomo, T., Nakata, M., and Kawabata, S. (2018). Competence-induced protein Ccs4 facilitates pneumococcal invasion into brain tissue and virulence in meningitis. *Virulence* 9, 1576–1587. doi: 10.1080/21505594.2018.1526530

Hobbs, J. K., Pluvinage, B., and Boraston, A. B. (2018). Glycan-metabolizing enzymes in microbe-host interactions: the *Streptococcus pneumoniae* paradigm. *FEBS Lett.* 592, 3865–3897. doi: 10.1002/1873-3468.13045

Ishiguro, T., Takayanagi, N., Yamaguchi, S., Yamakawa, H., Nakamoto, K., Takaku, Y., et al. (2013). Etiology and factors contributing to the severity and mortality of community-acquired pneumonia. *Intern. Med.* 52, 317–324. doi: 10.2169/internalmedicine.52.8830

Janapatla, R. P., Chen, C. L., Hsu, M. H., Liao, W. T., and Chiu, C. H. (2018). Immunization with pneumococcal neuraminidases NanA, NanB and NanC to generate neutralizing antibodies and to increase survival in mice. *J. Med. Microbiol.* 67, 709–723. doi: 10.1099/jmm.0.000724

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

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

Kawamura, Y., Hou, X. G., Sultana, F., Miura, H., and Ezaki, T. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *Int. J. Syst. Evol. Micrbiol.* 45, 406–408.

Kilian, M., and Tettelin, H. (2019). Identification of virulence-associated properties by comparative genome analysis of *Streptococcus pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, Three *S. oralis* Subspecies, and *S. infantis*. *mBio* 10:e01985-19. doi: 10.1128/mBio.01985-19

Kilian, M., Riley, D. R., Jensen, A., Bruggemann, H., and Tettelin, H. (2014). Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. *mBio* 5:e01490-14. doi: 10.1128/mBio.01490-14

King, S. J., Hippe, K. R., and Weiser, J. N. (2006). Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae*. *Mol. Microbiol.* 59, 961–974. doi: 10.1111/j.1365-2958.2005.04984.x

Limoli, D. H., Sladek, J. A., Fuller, L. A., Singh, A. K., and King, S. J. (2011). BgaA acts as an adhesin to mediate attachment of some pneumococcal strains to human epithelial cells. *Microbiology* 157(Pt 8), 2369–2381. doi: 10.1099/mic.0.045609-0

Lock, R. A., Hansman, D., and Paton, J. C. (1992). Comparative efficacy of autolysin and pneumolysin as immunogens protecting mice against infection by *Streptococcus pneumoniae*. *Microb. Pathog.* 12, 137–143. doi: 10.1016/0882-4010(92)90116-6

Lofling, J., Vimberg, V., Battig, P., and Henriques-Normark, B. (2011). Cellular interactions by LPxTG-anchored pneumococcal adhesins and their streptococcal homologues. *Cell Microbiol.* 13, 186–197. doi: 10.1111/j.1462-5822.2010.01560.x

Long, J. P., Tong, H. H., and DeMaria, T. F. (2004). Immunization with native or recombinant *Streptococcus pneumoniae* neuraminidase affords protection in the chinchilla otitis media model. *Infect. Immun.* 72, 4309–4313. doi: 10.1128/IAI.72.7.4309-4313.2004

Marion, C., Limoli, D. H., Bobulsky, G. S., Abraham, J. L., Burnaugh, A. M., and King, S. J. (2009). Identification of a pneumococcal glycosidase that modifies O-linked glycans. *Infect. Immun.* 77, 1389–1396. doi: 10.1128/IAI.01215-08

Mori, Y., Yamaguchi, M., Terao, Y., Hamada, S., Ooshima, T., and Kawabata, S. (2012). α-Enolase of *Streptococcus pneumoniae* induces formation of neutrophil extracellular traps. *J. Biol. Chem.* 287, 10472–10481. doi: 10.1074/jbc.M111.280321

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., and Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins* 12, 345–364. doi: 10.1002/prot.340120407

O'Neill, J. (2016). *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*. London: The Review on Antimicrobial Resistance.

Piao, Z., Akeda, Y., Takeuchi, D., Ishii, K. J., Ubukata, K., Briles, D. E., et al. (2014). Protective properties of a fusion pneumococcal surface protein A (PspA) vaccine against pneumococcal challenge by five different PspA clades in mice. *Vaccine* 32, 5607–5613. doi: 10.1016/j.vaccine.2014.07.108

Pond, S. L., Frost, S. D., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679. doi: 10.1093/bioinformatics/bti079

Radhakrishnan, D., Yamaguchi, M., Kawabata, S., and Ponnuraj, K. (2018). *Streptococcus pneumoniae* surface adhesin PfbA and its interaction with erythrocytes and hemoglobin. *Int. J. Biol. Macromol.* 120(Pt A), 135–143. doi: 10.1016/j.ijbiomac.2018.08.080

Rambaut, A. (2018). *FigTree ver.1.4.4.* Available: https://github.com/rambaut/figtree/releases. (accessed November 26, 2018).

Richards, V. P., Palmer, S. R., Pavinski Bitar, P. D., Qin, X., Weinstock, G. M., Highlander, S. K., et al. (2014). Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biol. Evol.* 6, 741–753. doi: 10.1093/gbe/evu048

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Salvadori, G., Junges, R., Morrison, D. A., and Petersen, F. C. (2019). Competence in *Streptococcus pneumoniae* and close commensal relatives: mechanisms and implications. *Front. Cell Infect. Microbiol.* 9:94. doi: 10.3389/fcimb.2019.00094

Sankar, S., Yamaguchi, M., Kawabata, S., and Ponnuraj, K. (2020). *Streptococcus pneumoniae* Surface Adhesin PfbA exhibits host specificity by binding to human serum albumin but not bovine, rabbit and porcine serum albumins. *Protein J.* 39, 1–9. doi: 10.1007/s10930-019-09875-y

Singh, A. K., Pluvinage, B., Higgins, M. A., Dalia, A. B., Woodiga, S. A., Flynn, M., et al. (2014). Unravelling the multiple functions of the architecturally intricate Streptococcus pneumoniae beta-galactosidase. *BgaA*. *PLoS Pathog.* 10:e1004364. doi: 10.1371/journal.ppat.1004364

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

Tanabe, A. S. (2008). *Phylogears2 ver. 2.0*. Available online at: http://www.fifthdimension.jp/ (accessed May 10, 2019).

Tanabe, A. S. (2011). Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. *Mol. Ecol. Resour.* 11, 914–921. doi: 10.1111/j.1755-0998.2011.03021.x

Tong, H. H., Li, D., Chen, S., Long, J. P., and DeMaria, T. F. (2005). Immunization with recombinant *Streptococcus pneumoniae* neuraminidase NanA protects chinchillas against nasopharyngeal colonization. *Infect. Immun.* 73, 7775–7778. doi: 10.1128/IAI.73.11.7775-7778.2005

Uchiyama, S., Carlin, A. F., Khosravi, A., Weiman, S., Banerjee, A., Quach, D., et al. (2009). The surface-anchored NanA protein promotes pneumococcal brain endothelial cell invasion. *J. Exp. Med.* 206, 1845–1852. doi: 10.1084/jem.20090386

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Yamaguchi, M. (2018). Synergistic findings from microbiological and evolutional analyses of virulence factors among pathogenic streptococcal species. *J. Oral Biosci.* 60, 36–40. doi: 10.1016/j.job.2018.02.004

Yamaguchi, M., Goto, K., Hirose, Y., Yamaguchi, Y., Sumitomo, T., Nakata, M., et al. (2019a). Identification of evolutionarily conserved virulence factor by selective pressure analysis of *Streptococcus pneumoniae*. *Commun. Biol.* 2:96. doi: 10.1038/s42003-019-0340-7

Yamaguchi, M., Hirose, Y., Takemura, M., Ono, M., Sumitomo, T., Nakata, M., et al. (2019b). *Streptococcus pneumoniae* evades host cell phagocytosis and limits host mortality through its cell wall anchoring protein PfbA. *Front. Cell Infect. Microbiol.* 9:301. doi: 10.3389/fcimb.2019.00301

Yamaguchi, M., Hirose, Y., Nakata, M., Uchiyama, S., Yamaguchi, Y., Goto, K., et al. (2016). Evolutionary inactivation of a sialidase in group B *Streptococcus*. *Sci. Rep.* 6:28852. doi: 10.1038/srep28852

Yamaguchi, M., Nakata, M., Sumioka, R., Hirose, Y., Wada, S., Akeda, Y., et al. (2017). Zinc metalloproteinase ZmpC suppresses experimental pneumococcal meningitis by inhibiting bacterial invasion of central nervous systems. *Virulence* 8, 1516–1524. doi: 10.1080/21505594.2017.1328333

Yamaguchi, M., Terao, Y., Mori, Y., Hamada, S., and Kawabata, S. (2008). PfbA, a novel plasmin- and fibronectin-binding protein of *Streptococcus pneumoniae*, contributes to fibronectin-dependent adhesion and antiphagocytosis. *J. Biol. Chem.* 283, 36272–36279. doi: 10.1074/jbc.M807087200

Yuan, Z. Q., Lv, Z. Y., Gan, H. Q., Xian, M., Zhang, K. X., Mai, J. Y., et al. (2011). Intranasal immunization with autolysin (LytA) in mice model induced protection against five prevalent *Streptococcus pneumoniae* serotypes in China. *Immunol. Res.* 51, 108–115. doi: 10.1007/s12026-011-8234-x

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.

Copyright © 2020 Yamaguchi, Takemura, Higashi, Goto, Hirose, Sumitomo, Nakata, Uzawa and Kawabata. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fmicb.2020.593548

Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2

Sheng Liu[1,2†], Jikui Shen[3†], Shuyi Fang[4], Kailing Li[4], Juli Liu[5], Lei Yang[5], Chang-Deng Hu[6,7] and Jun Wan[1,2,4,8]*

[1] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States, [2] Collaborative Core for Cancer Bioinformatics (C^3B) shared by Indiana University Simon Comprehensive Cancer Center and Purdue University Center for Cancer Research, Indianapolis, IN, United States, [3] The Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [4] Department of BioHealth Informatics, Indiana University School of Informatics and Computing, Indiana University – Purdue University Indianapolis, Indianapolis, IN, United States, [5] Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, United States, [6] Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN, United States, [7] Purdue University Center for Cancer Research, Purdue University, West Lafayette, IN, United States, [8] The Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, United States

OPEN ACCESS

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

Reviewed by:
Zhongming Zhao, University of Texas Health Science Center at Houston, United States
Renfeng Li, Virginia Commonwealth University, United States

***Correspondence:**
Jun Wan
junwan@iu.edu

[†] *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: *10 August 2020*
Accepted: *07 September 2020*
Published: *25 September 2020*

Citation:
Liu S, Shen J, Fang S, Li K, Liu J, Yang L, Hu C-D and Wan J (2020) Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2. Front. Microbiol. 11:593548. doi: 10.3389/fmicb.2020.593548

Four signature groups of frequently occurred single-nucleotide variants (SNVs) were identified in over twenty-eight thousand high-quality and high-coverage SARS-CoV-2 complete genome sequences, representing different viral strains. Some SNVs predominated but were mutually exclusively presented in patients from different countries and areas. These major SNV signatures exhibited distinguishable evolution patterns over time. A few hundred patients were detected with multiple viral strain-representing mutations simultaneously, which may stand for possible co-infection or potential homogenous recombination of SARS-CoV-2 in environment or within the viral host. Interestingly nucleotide substitutions among SARS-CoV-2 genomes tended to switch between bat RaTG13 coronavirus sequence and Wuhan-Hu-1 genome, indicating the higher genetic instability or tolerance of mutations on those sites or suggesting that major viral strains might exist between Wuhan-Hu-1 and RaTG13 coronavirus.

Keywords: COVID-19, SARS-CoV-2 genome, evolution, genetic variants, clustering, co-infection

INTRODUCTION

A novel betacoronavirus SARS-CoV-2 (Viruses CSGotICoTo, 2020) causing human coronavirus disease 2019 (COVID-19) was first reported in Wuhan, Hubei China in December 2019 (Wu et al., 2020a; Zhou et al., 2020; Zhu et al., 2020). The pandemic of SARS-CoV2 has infected more than 12 million people over 180 countries and areas around the world with a death over a half million as of July 9, 2020 (Dong et al., 2020). The most vulnerable group in this COVID-19 pandemic is elderly and those with different underlying medical conditions such as malnourish, hypertension, diabetes, cancer and cardiovascular abnormality (Guzik et al., 2020). Much effort has been devoted by scientists all over the world to understand the features of SARS-CoV2, particularly the viral genome variations. In some cases, viral genomic mutations play a key role in propagation of SARS-COV-2.

Viral mutation may alter the viral infectivity and pose an additional challenge for detection by the host cell, and thus it is critical to identify these mutations, especially in the context of vaccine design and drug development. Similar to other viruses, SARS-CoV-2 has been creating random mutations on the genome over time. Only some of mutations were caught and corrected by the virus's error correction machinery (Kupferschmidt and Cohen, 2020). Analysing these data can potentially monitor the viral transmission routes and identify novel mutations associated with the transmission (Zhang et al., 2020a). For example, Given 103 earlier genome sequence data, at least two clades of SARS-CoV-2 were found to be involved in the global transmission based on T > C mutation on a singleton site at 28144 of the complete genome, which was further termed as S clade (C28144) and L clade (T28144) (Tang et al., 2020). Evolutionary analyses suggested S clade appeared to be more related to coronaviruses in animals. Most recently, three major clusters of SNVs involved in the pandemic were found by comparing 160 SARS-CoV-2 genomes (Sanchez-Pacheco et al., 2020) with RaTG13 (Forster et al., 2020a). Researchers also employed standard phylogenomic approaches and compared consensus sequences representing the dominant virus lineage within each infected host (Forster et al., 2020a; Lai et al., 2020). Such information will be of important value for the development of vaccine, transmission monitoring and ultimately the control of the pandemic. However, most of these studies were based on limited numbers of SARS-CoV-2 genomes collected during early pandemic time, which might lead to debating conclusions (Chookajorn, 2020; Forster et al., 2020b; Kupferschmidt, 2020; Mavian et al., 2020; Rambaut et al., 2020; Sanchez-Pacheco et al., 2020). To date, more than 40,000 SARS-CoV-2 whole genome sequences have been uploaded to the online platform The Global Initiative on Sharing Avian Influenza Data (GISAID) database[1] (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). With the availability of increased sample size following SARS-CoV-2 spreading to almost all countries/areas in the world, it is feasible to provide a comprehensive and updated analysis of the viral genetic variations.

In this study, we took advantage of the mega-datasets collected by GISAID which published almost thirty thousand high-quality SARS-CoV-2 genomes with high coverage until June 15, 2020. Our comprehensive analyses clearly revealed distinct patterns of four major group mutations prominent in different countries and areas, suggesting representative SARS-CoV-2 strains correspondingly. We uncovered novel dynamic transmission and evolution patterns for groups of SARS-CoV-2 variants. A few hundred patients were found to have multiple groups of mutations simultaneously. Comparing with four bat coronavirus genomes, we found that alternations of nucleotides on SARS-CoV-2 genome tend to occur at the same sites where bat coronavirus sequences were different from Wuhan-Hu-1. Strikingly, some nucleotide substitutions on SARS-CoV-2 were apt to be the same as RaTG13 coronavirus sequences. We further investigated protein structure alternations caused by the amino acid (AA) changes due to high-frequent non-synonymous SNVs. Our novel genome-wide discoveries provided more detailed information and shed the light of studying SARS-CoV-2 which has been clouding over the world.

[1]https://www.gisaid.org/

MATERIALS AND METHODS

Collection of Sequences

Complete high-coverage coronavirus sequences were downloaded from GISAID database as of June 15, 2020. 28,212 coronavirus genomes isolated from humans and four bat Rhinolophus affinis were analyzed, including Bat CoVRaTG13 and RmYN02 from Yunnan Province, China, SL-CoVZC45, SL-CoVZXC21 from Zhejiang Province, China. White spaces within the sequences were removed. We aligned these sequences using minimap2 (Li, 2018) with the reference, the complete genome of Wuhan-Hu-1 (GISAID ID EPI_ISL_402125) by Wu et al. (2020a,b). The variants were annotated by ANNOVAR (Wang et al., 2010) using NCBI Reference Sequence: NC_045512.2.

Mutation Rate

Mutation rate is an important factor to monitor virus propagation and evolution (Zhao et al., 2004). In this study, we compared all other SARS-CoV-2 genome sequences against Wuhan-Hu-1 genome only. Hence we modified the formula (Zhao et al., 2004) to calculate mutation rate such that,

$$\mu = \frac{\sum_{i=1}^{N-1} \frac{m_i}{t_i}}{(N-1)\,L} \times 365$$

where μ is the mutation rate per site per year, N is the total genomes collected after Wuhan-Hu-1 with detailed time information, m_i is number of mutations of ith genome compared to Wuhan-Hu-1, t_i is the time difference in days when ith genome was collected after Wuhan-Hu-1, and L is the total length of SARS-CoV-2 genome (L = 29903).

Circos Plot

Circos plot (Krzywinski et al., 2009; Gu et al., 2014) was made given the ratios of genomes with SNV at each genome location of SARS-CoV-2. The concurrence ratio between two SNVs, X and Y, was defined as the ratio of the numbers of samples with both X and Y to the minimal number of samples with either X or Y.

$$\text{Concurrence ratio}\,(X, Y) = \frac{|X \cap Y|}{\min\left(|X|, |Y|\right)}$$

The connection lines in the Circos plot represent SNV pairs with high concurrence ratios (larger than 0.9).

Clustering of SNVs

Two-way clustering was performed to categorize the SNVs and samples with a distance function of one minus concurrence ratio on 54 frequent SNVs and about twenty-eight thousand samples.

Enrichment Analysis

While comparing SARS-CoV-2 genomic mutation sites and sites where Wuhan-Hu-1 varying from bats' coronavirus, we used

hypergeometric model to calculate the statistical significance of the overlaps.

Protein Structure Analysis

We used PyMOL (Schrödinger, Inc.) to visualize and analyze protein structure for WT (Wuhan-Hu-1) and mutations. Mutagenesis tools in PyMOL was utilized to detect if a clash was generated upon mutation. Properties of AAs were retrieved from the "Table of standard amino acid abbreviations and properties" on the Wikipedia[2]. The solved structures of Spike, nsp3, nsp5, nsp7 and Pol were downloaded from Protein Data Bank (PDB) (Berman et al., 2003): 6vyb for Spike using electron microscopy (Walls et al., 2020), 6w6y for nsp3 using X-ray diffraction method (Michalska et al., 2020), 6lu7 for nsp5 using X-ray diffraction method (Jin et al., 2020), 6wqd for nsp7 using X-ray diffraction method (Kim et al., 2020), 6m71 for Pol using electron microscopy (Gao et al., 2020). Structures of other proteins, e.g., nsp2, nsp4, nsp6, Hel, ExoN, ORF3a, M, ORF8, and N, were predicted by C-I-Tasser model[3] (Yang et al., 2015; Huang et al., 2020; Zhang et al., 2020b): QHD43415_2 for nsp2, QHD43415_4 for nsp4, QHD43415_6 for nsp6, QHD43415_12 for Hel, QHD43415_13 for ExoN, QHD43417 for ORF3a, QHD43419 for M, QHD43422 for ORF8, and QHD43423 for N.

RESULTS

Genetic Variants of SARS-CoV-2

We downloaded and analyzed 28,212 SARS-CoV-2 complete genome sequences after excluding low-coverage ones from the GISAID database. Using Wuhan-Hu-1 (NCBI Reference Sequence: NC_045512.2, GISAID ID: EPI_ISL_402125) as reference genome, we found that total 12,649 nucleotide sites had single nucleotide variants (SNVs) when compared to reference genome. The mutation rate was 1.1×10^{-3} per site per year for all point mutations, while it became 1.0×10^{-3} if we removed SNVs which appeared only once. Both sequence substitution rates were in the same range as $0.80–2.38 \times 10^{-3}$ for SARS-CoV genome as reported (Zhao et al., 2004). Majority of SNVs had very low occurrence frequency (**Figure 1A**), suggesting a high chance of random or unstable mutations. Four nucleotide substitutions were detected in over 70% of genome sequences: A23403G, C3037T, C14408T, and C241T. They distributed at distinct SARS-CoV-2 genome locations, on the gene body of Spike, ORF1a, and ORF1ab, and upstream of ORF1ab, respectively. Additionally, there were other 50 unique SNVs arose from larger than 1% of populations ($n > 282$). It is interesting that some of these frequent SNVs occurred almost simultaneously with concurrent ratio larger than 0.9 (see methods) as shown by blue-line connections in **Figure 1A**. They may appear across different proteins. For example, A23403G changes an aspartate to a glycine on Spike (D614G) while C14408T converts a proline to a leucine on ORF1ab (P4715L). Over 99% of both SNVs were found simultaneously on more than 74% samples, suggesting a biological connection between these concurrent variant sites.

Among total 54 frequent SNVs, 31 mutations are non-synonymous variants or non-frameshift substitutions (**Figure 1B**). Some of them have been discussed separately by previous studies (Forster et al., 2020a; Guan et al., 2020; Tang et al., 2020) or marked as elements in clades G, S, and V from the GISAID report (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). Here, two-way clustering was performed on 54 frequent SNVs and about 28,000 samples (**Figure 1B**). It is clear to see four major groups of SNVs covering almost all samples, including groups A (C14408T/A23403G, occurring on 21,116 samples), B (T28144C on 2,802 samples), C (G11083T/G26144T on 3,173 samples), and D (G1440A/G2891A on 441 samples). Most SNVs belonged to one unique cluster, while a few SNVs crossed different groups. Taken as an example, a synonymous mutation C14805T existed in both group B and C (**Figure 1B**), covering over 8% of worldwide samples. Majority (79%) of C14805T can be another signature mutation in group C with SNVs G11083T and G26144T together. In general, the geographical locations of infected patients bearing these special groups of mutations were very different.

Forty countries and areas with numbers of viral genomes larger than 50 were chosen to probe the geographical distributions of these SNVs. Group A, represented by two non-synonymous mutations, A23403G and/or C14408T, was borne in totally 72% of samples in the study, including about 82% from Europe and 67% from North America (**Figure 1B**). The top three countries with the highest ratio in group A (**Figure 2A**) were Russia (99%), Denmark (96%), and South Africa (96%). Group B was distinguished by non-synonymous mutation T28144C (**Figure 1B**) which results in substitution of a leucine by a serine on ORF8. It was projected in Thailand (50%), Spain (41%), China (31%), and some other Asian countries/areas (**Figure 2B**). Group C was featured by two non-synonymous SNVs, G11083T and G26144T (**Figure 1B**), which substituted a leucine with a phenylalanine on ORF1ab and a glycine with a valine on ORF3a, respectively. This group existed in many Asian and European countries/areas, e.g., Hong Kong, Singapore, Japan, Turkey etc. (**Figure 2C**), as reported previously (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Forster et al., 2020a; Guan et al., 2020; Tang et al., 2020). Group D includes two non-synonymous SNVs, G1440A and G2891A (**Figure 1B**), both of which change the AA sequences on ORF1ab. It confirms the clade D, previously defined by Guan et al. (2020) based on smaller set of patients. G1440A led to the AA change, G212D on non-structural protein 2 (nsp2), while G2891A caused A58T on non-structural protein 3 (nsp3). D-group was mainly found in several European countries/areas, e.g., Wales (17%), Germany (10%), and Belgium (5%) (**Figure 2D**).

Besides signature variants in each major group discussed above, some SNVs were found in relatively smaller populations but concurred with the major signature SNVs. Importantly, many non-major SNVs were mutually exclusively presented with each other in different countries and areas (**Figure 1B**). For instance, about 28 mutations coinciding with A23403G and C14408T in the group A composed sub-types of A (**Figure 1B**), e.g., G25563T

[2]https://en.wikipedia.org/wiki/Amino_acid

[3]https://zhanglab.ccmb.med.umich.edu/COVID-19/

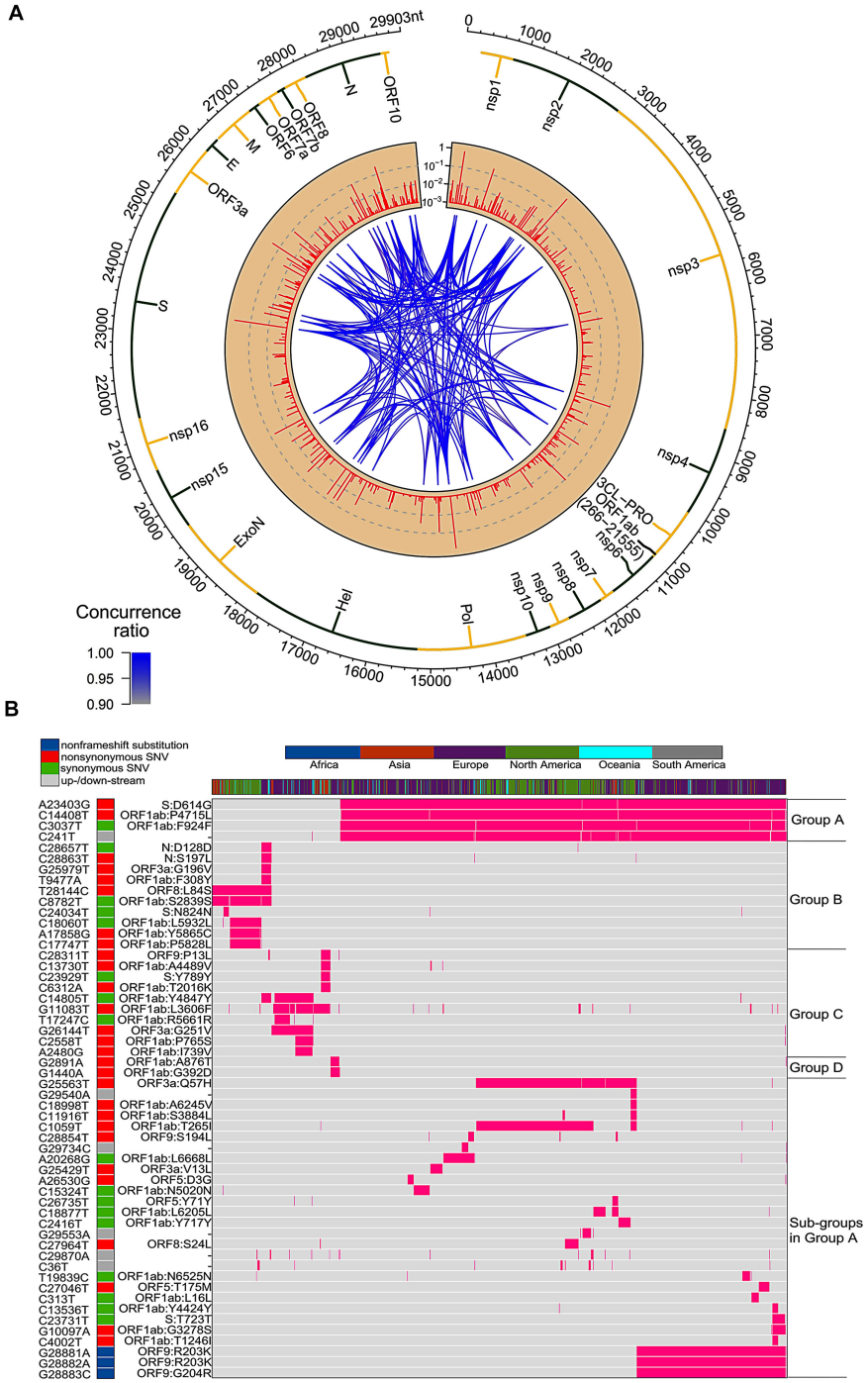


FIGURE 1 | SNVs on about thirty thousand SARS-CoV-2 complete genomes. **(A)** Circos plot shows distribution, frequency, and co-occurrences of SNV2. From outer to inner circle: coronavirus genome location (nt), gene annotation, occurrence ratios of SNVs at the site (log10 scale, red bars), and connections with high concurrence rates (>0.9) represented by blue lines. The darker the blue lines, the higher concurrence rates. **(B)** Fifty-four high frequent SNVs with annotated AA changes were detected (in purple) in about twenty-eight thousand patients worldwide. Four major clusters of SNVs and consequent subgroups can be formed to represent patients from different geographical locations.

and C1059T. However, two separable sets of samples were associated with different combinations of G25563T and C1059T (**Figure 2A**). Sub-cluster A1 included both G25563T and C1059T, whereas sub-cluster A2 had G25563T but excluded C1059T. They may represent divergent strains found in distinct populations from varied countries/areas. Specifically, A1 occurred in 21%

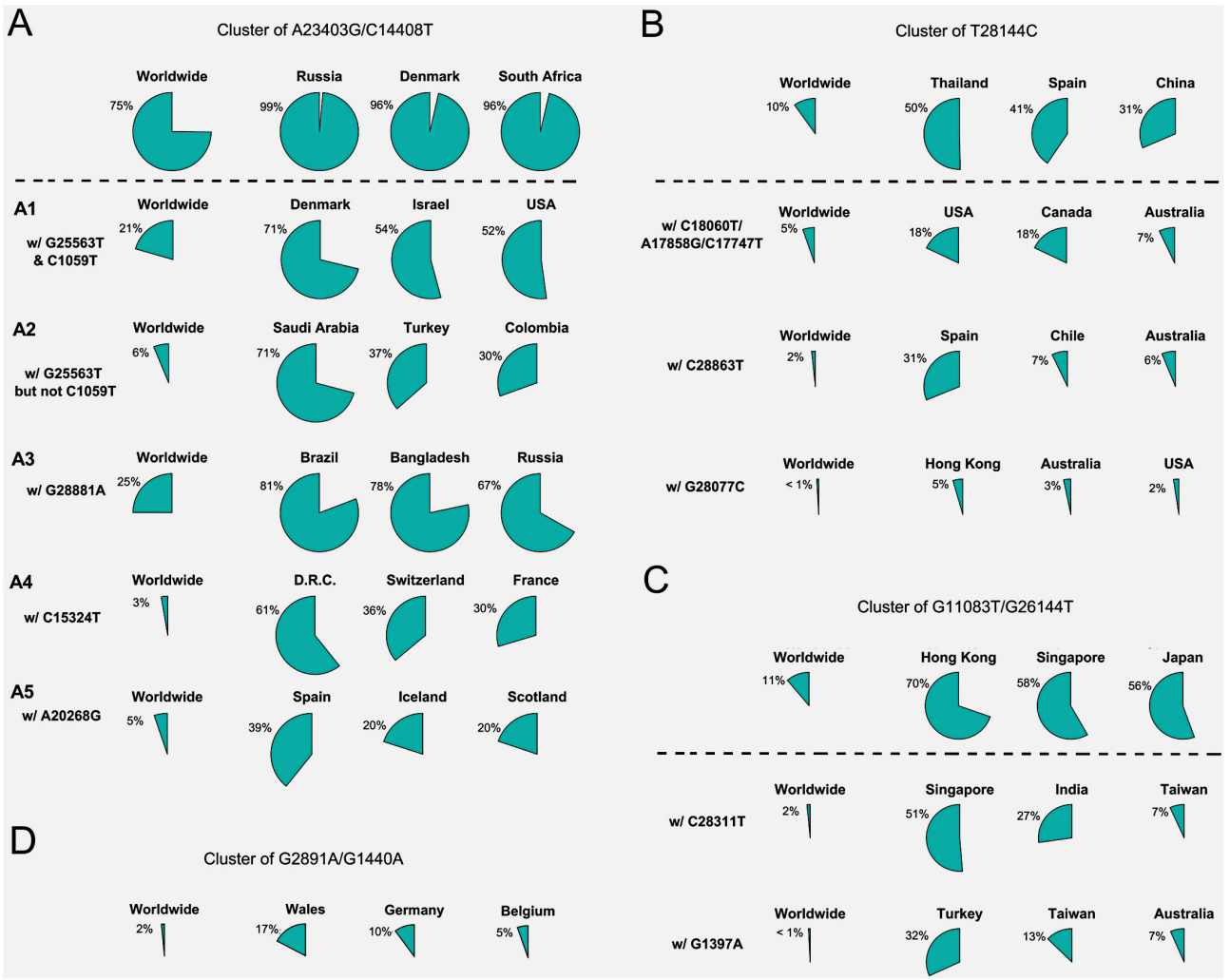


FIGURE 2 | Frequencies of signature SNVs in worldwide and top three countries/areas. **(A–D)** Four major groups and/or their sub-groups had distinct representing countries/areas, indicating different transmission sources and evolution paths.

of all SAR-CoV-2 genomes collected world widely, particularly in 71% of Denmark, 54% of Israel, and 52% of United States, whereas A2 was found in only 6% population, which were mostly discovered in Saudi Arabia (71%), Turkey (37%), and Columbia (30%). Another sub-cluster, A3, had consecutive mutations at positions 28881-28883 on SARS-CoV-2 complete genome, leading non-frameshift substitutions on ORF9: R203K-G204R. A3 occupied 25% of worldwide cases, represented by three major countries, Brazil (81%), Bangladesh (78%), and Russia (67%). Even though some sub-clusters of mutations were found in smaller worldwide populations (around or lesser than 5%), they were significantly over-represented in several countries and areas. For instance, A4 with synonymous mutation C15324T was detected in 61% samples of an African country, Democratic Republic of the Congo (DRC), coming together with 36% of Switzerland and 30% of France.

Patients from one country may have different main groups or sub-types of mutations. A synonymous A20268G in cluster A5 (**Figure 2A**) was sampled in Spain (39%), Iceland (20%), and Scotland (20%). It is interesting that other 41% of Spain samples had another distinguished non-synonymous mutation T28144C in group B (**Figure 2B**), same as many samples from Asian patients. It suggests the viral transmission path on these patients. 31% of Spain samples also had another unique mutation, C28863T, substituting a serine with a leucine on ORF9, concurrent with T28144C. About 18% of Australia samples were found in group B as well. But they came with additional diverse mutually exclusive SNVs, e.g., either C18060T/A17858G/C17747T (7%), or C28863T (6%), or G28077C (3%) as shown in **Figure 2B**. Similar scenarios were observed in United States, where approximately 18% of samples encompassed T28144C with C18060T/A17858G/C17747T, while another 2% was recognized with a different non-synonymous mutation G28077C in the same main group B (**Figure 2B**).

SNVs in group C including G11083T and G26144T existed in many Asian countries and areas (**Figure 2C**), such as Hong Kong,

Singapore, Japan, Indian, Taiwan, and Turkey, as reported previously (Forster et al., 2020a). However, different countries and areas were distinguished by extra variants in the same prime group C. For example, 51% of Singapore was detected with non-synonymous C28311T on ORF9, while Turkey had 32% samples with non-synonymous G1397A on ORF1ab.

Currently, it lacks sufficient evidences to make a conclusive statement about the origins of all SARS-CoV-2 mutations. But time-annotated data collections can still explore geographical evolution patterning of specific SNVs, albeit limited number of high-quality and high-coverage sequenced viral genomes at some time points. For example, only three cases with mutations T28144C and C18060T (one sub-type in group B) reported in Washington State of United States in January 2020, in addition to eight cases in China and additional one in Singapore at almost the same time (**Figure 3A**). It is notable that T28144C and C18060T concurred with additional *de novo* non-synonymous mutations C17747T/A17858G on ORF1ab in 51 cases from United States and another one from the cruise of Grand Princess in February 2020. No such case was detected in other countries/areas. One month later, this group of signature variants spread over many states of United States, particularly west coast of United States, and other countries and areas of different continents, including Canada, Australia, Iceland, Mexico, New Zealand, and England etc.

Over half of American patients had been sampled with mutations C14408T/A23403G and C1059T on SARS-CoV-2 genome (**Figure 2A**). Retrieving data in January 2020, we found only one case with both C14408T and A23403G in China from our dataset (**Figure 3B**). The first case in United States was reported in New Hampshire at the east coast concurrently with C1059T, in addition to five in France, one in Belgium and one in Senegal. The numbers of such cases boosted up in United States and other countries/areas in March 2020, including 354 in Denmark, 164 in Australia, and 98 in France, 77 in England, etc. In the United States, approximately 1,000 cases were found on the east coast of United States, while over 400 cases were identified on the west coast as well.

The variants of group A (C14408T/A23403G) indicated at least two strains of SARS-CoV-2 distinguishable on the sites of Spike and ORF1ab. One viral strain observed from Wuhan-Hu-1 can be named as DP with an aspartate on 614 of Spike and a proline on 4715 of ORF1ab, while another potential one, named as GL, had a glycine on the site of 614 on Spike and a leucine on the site of 4715 on ORF1ab instead. The ratio of GL strain in all United States cases increased dramatically from 6% in February to 87% in May and June 2020 (**Figure 3C**). The similar growing trend was observed in most of other countries, regardless when this group of mutations were first present (**Figure 3C**). In general, 91% of samples from all these countries had strain GL since May 2020 compared to only 3% in February (**Figure 3D**), suggesting that the GL strain of SARS-CoV-2 might become much more stable and prevailing than the other strain DP like Wuhan-Hu-1 after 6-month evolution and transmission.

Different groups of mutations also exhibited distinguished evolution patterns (**Figures 3C,D**). Taking B-group SNVs for example, we found that the ratio decreased over time from 35% in January 2020 to almost zero in June in these countries, indicating that at least two strains existed at the early of COVID-19 pandemic. However, strains including variant at 28,144 other than Wuhan-Hu-1 almost diminished after 7 months of transmission. Only the strain that has the same nucleotide T28144 as Wuhan-Hu-1 finally became the most stabilized strain in the host. The similar patterns were observed for groups C and D as well, even though a sudden increasing was found in February and/or March 2020 due to unknown reasons. For instance, in group C with SNVs G1440A and G2891A, Germany had a high ratio, 47.8% (11 out of 23), in February, while 25% of (96 out of 384) Wales were sampled with the same variants in March 2020.

Four main groups of mutations showed mutually exclusive in about twenty-eight thousand patients, indicating at least five unique viral strains (including the one same as Wuhan-Hu-1) potentially existed in the host. However, as reported in early of March 2020, a patient hospitalized in Iceland infected by two SARS-CoV-2 subtypes simultaneously[4]. One strain of the SARS-CoV-2 coronavirus was more aggressive, according to Reykjavik Grapevine newspaper citing CEO of CODE Genetics biopharmaceutical company Kari Stefansson. The second strain is a mutation from the original version of the coronavirus that appeared in Wuhan, China. This was regarded as the first known case of co-infection. Gudbjartsson et al. (2020) reported that patient T25 carries both the A2a1a strain and the A2a1a + 25958 strain. As shown in **Figure 4**, we found that 13 genomes bore both A-group and B-group mutations, while 347 genomes had variant groups of A and C, and both B and C groups were involved in 44 genomes. Strikingly, one patient from Spain was detected with three groups of variants simultaneously, A–C. 17 and 4 out of 441 SARS-CoV-2 genomes with D-group SNVs were overlapped with groups A and C, respectively.

Comparison of Variants Between SARS-CoV-2 Genomes and Bat Coronavirus Sequences

Bats were regarded as reservoir species for SARS-CoV-2. To understand potential associations between SNVs among SARS-CoV-2 genomes from patients and bat coronavirus sequences, we also aligned four bat coronavirus sequences to Wuhan-Hu-1 complete genome. The ratios of variants between Wuhan-Hu-1 and bats were 3.8% (RaTG13), 11.1% (bat-SL-CoVZC45), 11.1% (bat-SL-CoVZXC21), and 4.8% (RmYN02). As described above, 12,649 out of 29,903 nts (42.3%) on SARS-CoV-2 genome underwent variation among about 28,000 samples. Interestingly, the ratios of SARS-CoV-2 SNVs on the sites where bats' sequences differed from Wuhan-Hu-1 were significantly elevated (**Figure 5A**). Among them, RaTG13 reached the highest ratio (61.5%) with $p = 2.7e\text{-}40$. The result suggests that the sites where Wuhan-Hu-1 differed from bats might have higher tolerance for sequence variations or higher genetic instability.

In theory, 12,649 identified SARS-CoV-2 SNVs can potentially turn to be any one of three nucleotides other than the original ones from Wuhan-Hu-1. When we focused on the sites where

[4]https://www.mbs.news/a/2020/03/icelandic-man-reportedly-caught-two-coronavirus-subtypes-simultaneously.html

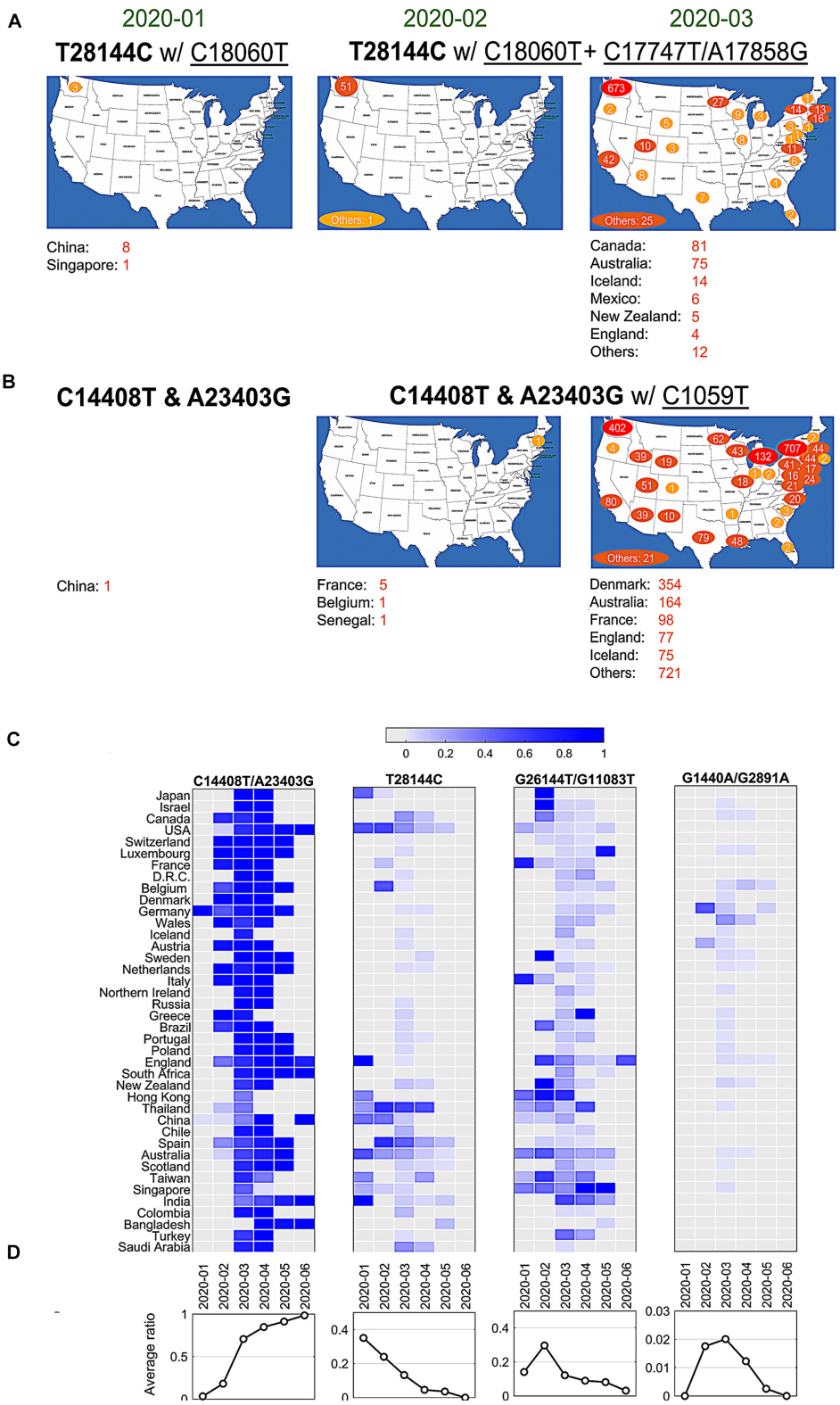


FIGURE 3 | SNVs migration and evolution patterns over time. **(A)** SNV T28144C with C18060T and additional C17747T/A17858G spread in United States and other countries/areas from January to March of 2020. **(B)** SNVs of C14408T and A23403G with C1059T spread in United States and other countries/areas from January to March of 2020. **(C)** The ratios of four significant groups of SNVs, **(A–D)** in **Figures 1** and **2**, varied in different countries/areas with time development. **(D)** Average temporal ratios of groups **(A–D)** SNVs show distinct patterns from January to June of 2020.

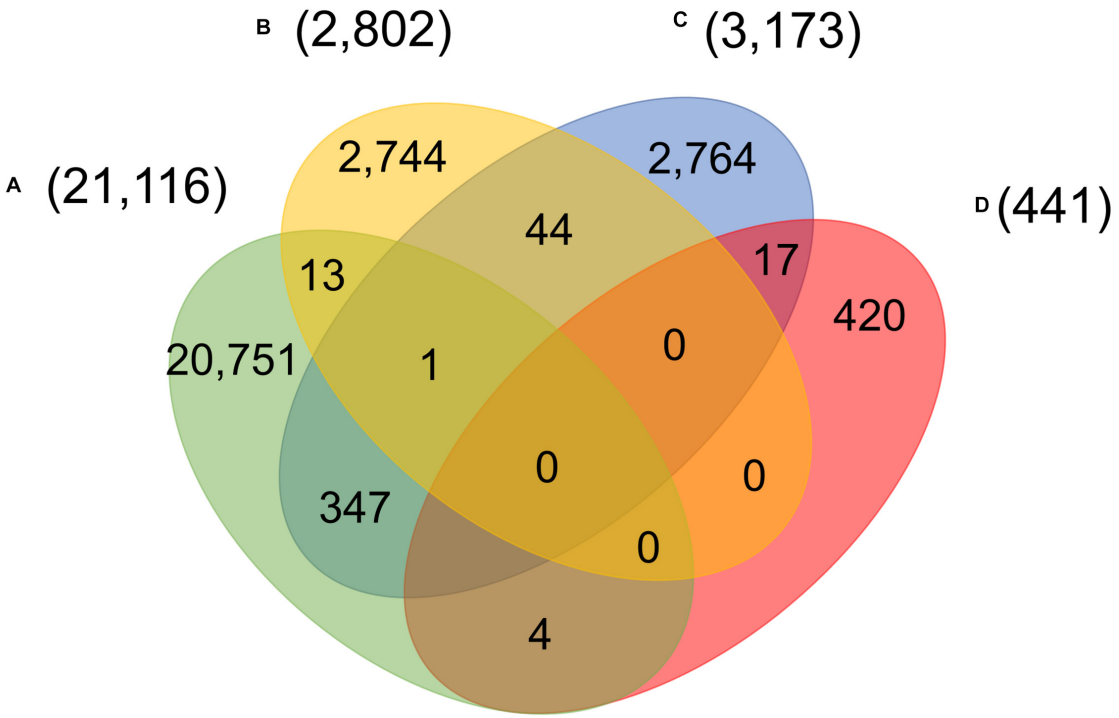


FIGURE 4 | Number of patients carrying significant groups **(A–D)** of SNVs, indicating potential co-infection by different SARS-CoV-2 strains.

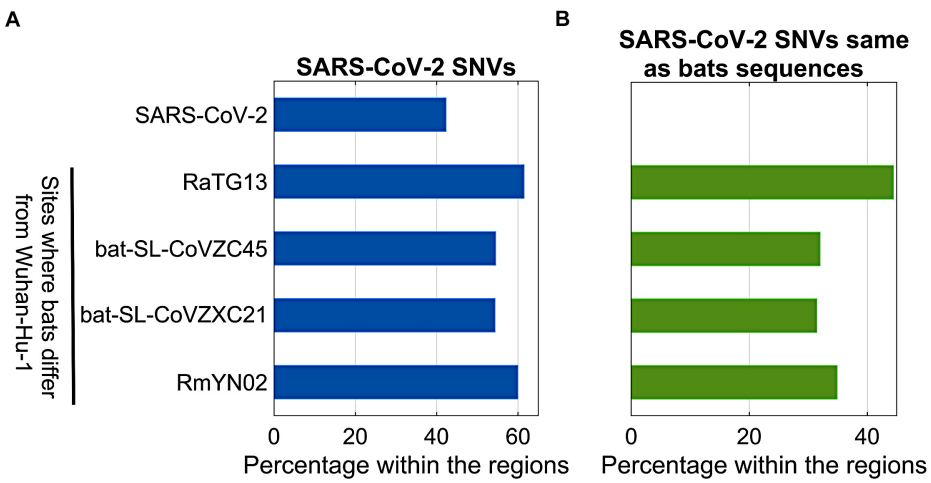


FIGURE 5 | Comparison between SARS-CoV-2 SNVs and nucleotides of bats (RaTG13, bat-SL-CoVZC45, bat-SL-CoVZXC21, and RmYN02) coronavirus varying from Wuhan-Hu-1. **(A)** Percentage of SARS-CoV-2 SNVs on different regions, including SARS-CoV-2 complete genome, and sites where bats coronavirus differ from Wuhan-Hu-1; **(B)** Percentage of SARS-CoV-2 SNVs which converted to bats nucleotides within the same regions as shown in **(A)**.

bats coronavirus sequences differed from Wuhan-Hu-1, it turned out that SARS-CoV-2 SNVs had the same mutated nucleotides as RaTG13 coronavirus does on 503 out of 1,132 (44.4%) sites where RaTG13 coronavirus sequence differed from Wuhan-Hu-1 (**Figure 5B**), including C29095T (Forster et al., 2020a) and seven high frequent SNVs identified from our major groups, e.g., C2416T and C3037T from group A, C8782T, C18060T, C24034T, and T28144C from group B, and C23929T in group C. The ratio for RaTG13 coronavirus was much higher than the ratios observed in other three bat coronavirus sequences (32.0, 31.4, and 34.9%, respectively).

SARS-CoV-2 SNVs and Protein Functions

Viral sequence mutations will likely affect viral infection, replication, and/or propagation, and thus alter SARS-CoV-2 transmission properties and COVID-19 severity. As previous

reported, S68F and P71L non-synonymous mutation in E-protein of SARS CoV-2 were the most common mutation in in E-protein (Hassan et al., 2020); Q57H, G251V, and G196V non-synonymous mutation in ORF3a of SARS-CoV-2 would link to the virulence, infectivity, ion channel activity and viral release (Issa et al., 2020); deletion of ORF8 leads to increased production of the interferon and reduced level of inflammatory cytokines (Chen et al., 2020; Gong et al., 2020; Li et al., 2020a). Most recently, researchers found that D614G non-synonymous mutation located in spike protein would increase infectivity (Eaaswarkhanth et al., 2020; Li et al., 2020b; Yurkovetskiy et al., 2020; Zhang et al., 2020c).

Here, we analyzed the structure changes of 31 high-frequent non-synonymous SNVs and non-frameshift substitutions (**Figure 1B**) using PyMOL (Schrödinger, Inc.). It was interesting that all of them were on the surface area of corresponding proteins (**Table 1**). Eight of them clashed with nearby AAs on non-structural protein 2 (nsp2): G212D, nsp3: A58T, nsp4: F308Y, nsp5: G15S, ORF3a: V13L and G251V, ORF9: G204R, and Pol: A97V, which might be worthy of further analysis. Six of SNVs on S: D614G, ORF3a: Q57H, ORF5: D3G, ORF9: G204R, nsp2: G212D, nsp3: T1198K changed the charge upon the mutations. These changes may contribute to transmission and virulence of SARS-CoV-2. For example, a non-synonymous SNV, G1440A in group D, discovered in over 400 samples from several European countries, led to the AA change of G212D on nsp2. Such a change of G212D may add clashes between residues 212 and ASN183 (**Figures 6A,B**). It is interesting that Nsp2 G212D falls on the region homologous to the endosome-associated protein similar to the avian infectious bronchitis virus (PDB 3ld1), which plays a key role in the viral pathogenicity (Angeletti et al., 2020).

Similar to SARS coronavirus, SARS-CoV-2 enters into human cells through the binding of its receptor-binding domain (RBD) on Spike protein to human receptor ACE2 (Hoffmann et al., 2020b). Furin is responsible for the proteolytic cleavage of Spike protein to facilitate virus entry. In SARS-CoV-2, 15-nt CCTCGGCGGGCACGT encodes five AAs: PRRAR (681-685), locating at 23603–23617 of Wuhan-Hu-1 complete genome. Furin cleavage site bears a RXXR pattern (Molloy et al., 1992; Shiryaev et al., 2013). R685 makes an ideal furin proteolytic cleavage site (Coutard et al., 2020). Out of almost thirty thousand samples, 58 SARS-CoV-2 genomes were detected with mutations in the region, including 13 from England and 23 from United States. Non-synonymous SNVs, C23604T, was most frequent among others, causing the mutation of P681L. Other AA mutations included P681H/L/S, R682Q/W, R683P/Q, and A684E/T/S/V. As described previously, D614G on Spike caused by the SNV A23403G in group A covered about three-quarter of total sequenced genomes in our study. D614G on Spike did not generate clashes from the protein structure predictions (**Figures 6C,D**). However, the residue variations changed the negative polar side chain to neutral non-polar side chain (**Figures 6C,D** and **Table 1**). Since the site is close to the furin region, such alternation might be able to affect the interactions between furin and furin cleavage sites, then further influence cell-cell fusion and ability to infection (Hoffmann et al., 2020a).

In addition, Cryo-EM-based structural analysis (Wrapp et al., 2020) revealed that 5 key AAs within 434–507 of Spike protein contributed most to the binding activity. This was also confirmed by several recent cryo-EM structural studies (Shang et al., 2020a,b; Wang et al., 2020a; Yuan et al., 2020). The key AAs of SARS-CoV-2 RBD are: L455, G482, V483, E484, G485, F486, Q493, S494, and N501. Interestingly, we identified several non-synonymous SNVs of SARS-CoV-2 on L455, V483, G485, and S494 from sequenced sample, for instance, G22927T (L455F), G23009T (V483F), T23010C (V483A), G23105A (G485S), and T23042C (S494P). Among them, 28 viral genomes had mutation T23010C (V483A), all of which were sampled in United States, including 26 from Washington State. The RBD for SARS-CoV-2 has residues and motifs found in all three clades in lineage B of betacoronavirus (Letko et al., 2020), suggesting distinct cell entry pattern than that of other clades. L455, G485, F486, and N501 are among contact points of virus to human ACE2, changes in these positions may affect the strength of transmission of the virus.

DISCUSSION

In this study, we comprehensively analyzed almost thirty thousand high-quality and high-coverage SARS-CoV-2 complete genome sequences as well as four bat genome sequences. Even though some SNVs were reported previously and discussed individually, we used bioinformatics approaches to systematically identify four major mutually exclusive groups of SNVs among all samples, suggesting at least five viral strains existing (including one strain same as the reference). These mutations were detected in populations from different geographical locations. The results could provide some insights of possible new functions of SARS-CoV-2 proteins and further bring therapeutic potentials.

Distinct time-course evolution patterns were observed for four major groups of mutations. Some viral strains, e.g., GL with mutations C14408T and A23403G, may gradually replace Wuhan-Hu-1, to become dominant after several month evolutions. Or others may be eliminated naturally with time development, e.g., strains associated with groups B–D mutations (**Figures 3C,D**). It is hard to explain aberrant emergence of some strains, e.g., the peak time of groups C and D in February and March (**Figures 3C,D**), particularly due to the lack of enough numbers of high-quality sequenced samples worldwide, including China and other countries/areas, before February 2020. However, with more and more clinical data generated, evolution patterning associated with specific biological functions may be clearly uncovered. For example, several groups recently reported that A23403G mutation in Spike protein might alter the antigenic property and transmission ability due to the change of Spike-ACE2 interaction (Becerra-Flores and Cardozo, 2020; Korber et al., 2020a).

In general, four SNV clusters were mutually exclusively presented. But we still noticed a few hundred patients who were identified to carry multiple groups of SNVs simultaneously. Without clear evidence that homologous recombination in these regions in the intermediate or human host could occur

TABLE 1 | Protein structures changes corresponding to high frequent non-synonymous SNVs identified.

Protein	SNV	REF Charge	REF Polar	REF Molar mass	ALT Charge	ALT Polar	ALT Molar mass	Side of surface	Predict clash	Size change in molar Mass	Charge change	Polar change	Equivalent protein	Equivalent protein SNV
S	D614G	Negative	Polar	133	Neutral	Non-polar	75	Outer		−57	Yes	Yes		
ORF3a	G251V	Neutral	Non-polar	75	Neutral	Non-polar	117	Outer	Leu219	42	No	No		
ORF3a	G196V	Neutral	Non-polar	75	Neutral	Non-polar	117	Outer		42	No	No		
ORF3a	Q57H	Neutral	Polar	146	Positive	Polar	155	Outer		9	Yes	No		
ORF3a	V13L	Neutral	Non-polar	117	Neutral	Non-polar	131	Outer	Val80	14	No	No		
ORF5/M	D3G	Negative	Polar	133	Neutral	Non-polar	75	Outer		−57	Yes	Yes		
ORF5/M	T175M	Neutral	Polar	119	Neutral	Non-polar	149	Outer		30	No	Yes		
ORF8	S24L	Neutral	Polar	105	Neutral	Non-polar	131	Outer		26	No	Yes		
ORF8	L84S	Neutral	Non-polar	131	Neutral	polar	105	Outer		−26	No	Yes		
ORF9/N	P13L	Neutral	Non-polar	115	Neutral	Non-polar	131	Outer		16	No	No		
ORF9/N	S194L	Neutral	Polar	105	Neutral	Non-polar	131	Outer		26	No	Yes		
ORF9/N	S197L	Neutral	Polar	105	Neutral	Non-polar	131	Outer		26	No	Yes		
ORF9/N	R203K	positive	Polar	174	positive	Polar	146	Outer		−28	No	No		
ORF9/N	G204R	Neutral	Non-polar	75	positive	Polar	174	Inner	Met411	100	Yes	Yes		
ORF1ab	T265I	Neutral	Polar	119	Neutral	Non-polar	131	Outer		12	No	Yes	Nsp2	T85I
ORF1ab	G392D	Neutral	Non-polar	75	Negative	Polar	133	Outer	Asn183	57	Yes	Yes	Nsp2	G212D
ORF1ab	I739V	Neutral	Non-polar	131	Neutral	Non-polar	117	Inner		−14	No	No	Nsp2	I559V
ORF1ab	P765S	Neutral	Non-polar	115	Neutral	Polar	105	Outer		−10	No	Yes	Nsp2	P585S
ORF1ab	A876T	Neutral	Non-polar	89	Neutral	Polar	119	Outer	Ile62	30	No	Yes	Nsp3	A58T
ORF1ab	T1246I	Neutral	Polar	119	Neutral	Non-polar	131	Inner		12	No	Yes	Nsp3	T428I
ORF1ab	T2016K	Neutral	Polar	119	positive	Polar	146	Outer		28	Yes	No	Nsp3	T1198K
ORF1ab	F3071Y	Neutral	Non-polar	165	Neutral	Polar	181	Outer	Phe71	16	No	Yes	Nsp4	F308Y
ORF1ab	G3278S	Neutral	Non-polar	75	Neutral	Polar	105	Outer	Lys97	30	No	Yes	Nsp5	G15S
ORF1ab	L3606F	Neutral	Non-polar	131	Neutral	Non-polar	165	Outer		34	No	No	Nsp6	L37F
ORF1ab	S3884L	Neutral	Polar	105	Neutral	Non-polar	131	Outer		26	No	Yes	Nsp7	S25L
ORF1ab	A4489V	Neutral	Non-polar	89	Neutral	Non-polar	117	Outer	Gln117	28	No	No	Pol/nsp12	A97V
ORF1ab	P4715L	Neutral	Non-polar	115	Neutral	Non-polar	131	Outer		16	No	No	Pol/nsp12	P323L
ORF1ab	P5828L	Neutral	Non-polar	115	Neutral	Non-polar	131	Outer		16	No	No	Hel/nsp13	P504L
ORF1ab	Y5865C	Neutral	Polar	181	Neutral	Polar	121	Outer		−60	No	No	Hel/nsp13	Y541C
ORF1ab	A6245V	Neutral	Non-polar	89	Neutral	Non-polar	117	Outer		28	No	No	ExoN/nsp14	A320V

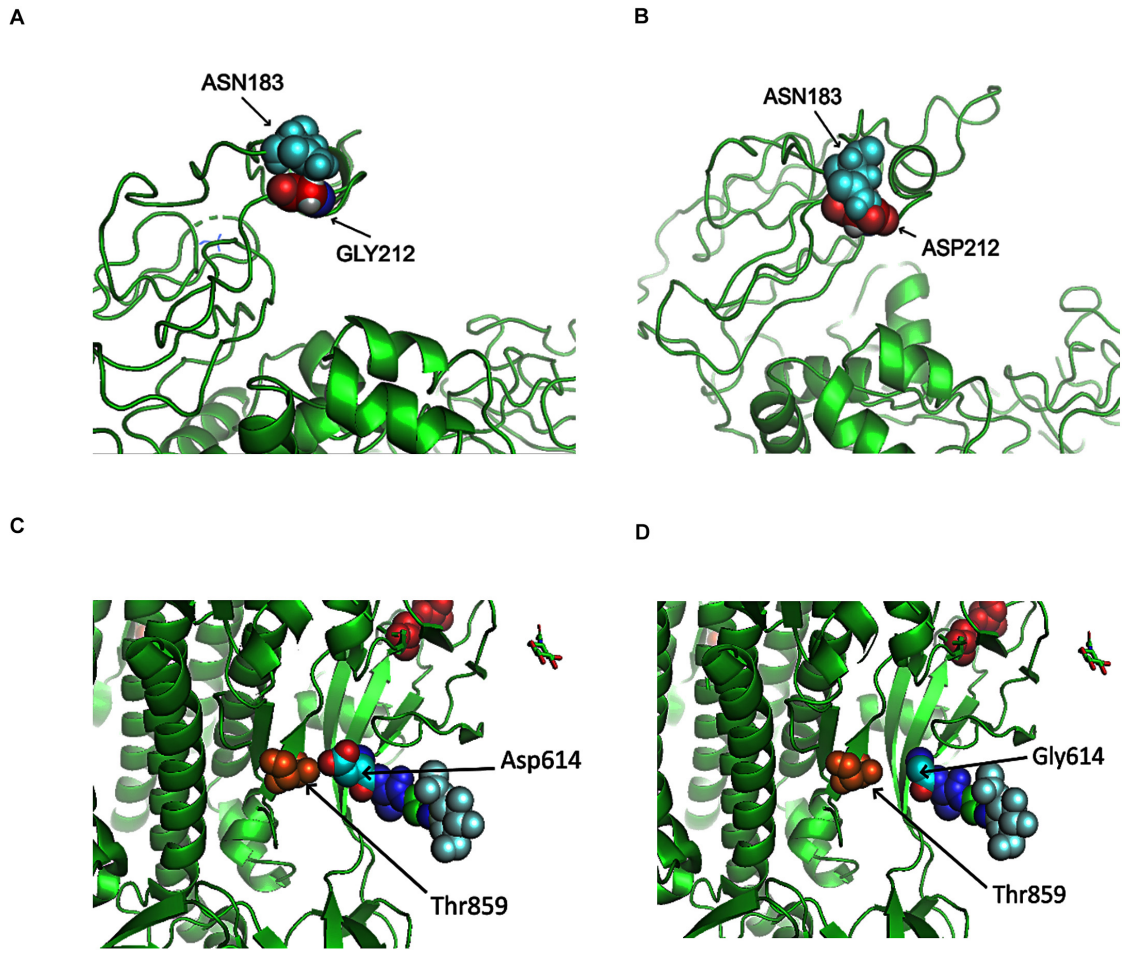


FIGURE 6 | Structures of SARS-CoV-2 non-structural protein 2 (nsp2) and Spike near specific mutation sites. **(A)** No clash was predicted between nsp2 GLY212 and nsp2 ASN183 in Wuhan-Hu-1. **(B)** A clash was predicted between ASP212 and ASN183 after the mutation G212D. **(C)** There was interaction between T859 and D614. **(D)** No contact was precited between T859 and G614 after the mutation D614G.

in these viruses, we just defined such overlaps as potential co-infections based on our observations and current knowledge. One possibility is that two or three strains co-existed and prevailed in the population of the same region during the periods when the patients got infected from other people. The patients could be first infected with one strain then another one later, suggesting that primary infection did not yield immunity in time against the subsequent infection from a different strain. Another possibility is that the virus underwent mutations during the transmission to another human due to the special environment of the host, consequently multiple representative mutations were present on the same patient. There are several other scenarios in addition to co-infections, including doubtful sequencing errors or cross sample contaminations. Unfortunately, it lacks of enough information at this moment about the potential post-infection immunity that has important implications for the epidemiologic assessment for the transmission (Kirkcaldy et al., 2020). Of course, the percentage of co-infection cases was less than 1.5% in this study. It might be the consequences of the quarantine and lockdown policy enforced after the spread of COVID-19, while social distancing and wearing face mask are considered effective approaches in reducing the chance of co-infections (Cheng et al., 2020; Eikenberry et al., 2020; Wang et al., 2020b; West et al., 2020). These policies reduced the likelihood that people met patients with different SARS-CoV-2 strains at the same time.

We further compared SNVs among SARS-CoV-2 genomes from human patients to bat coronavirus sequences. It is interesting that SARS-CoV-2 SNVs, particularly those high-frequent mutations, tend to occur at the same sites where bats coronavirus sequences varied from Wuhan-Hu-1, suggesting the high tolerance of these sites for genetic mutations, or potentials of SARS-CoV-2 turning to a wild-type pathogenic phenotype. RaTG13 coronavirus was most similar to SARS-CoV-2 from perspective of sequences, but it held the highest ratio of SARS-CoV-2 variants which converted to the bat's coronavirus sequences at the same sites. This suggests that

some strains of SARS-CoV-2 deviated from Wuhan-Hu-1 might be more similar to bat coronavirus RaTG13 than other bat coronavirus strains presented in this paper. Of course, we don't have more evidence to show the exact connections between them, but our results may shed the light to search intermediate host and further understand the mechanisms of interspecies transmission in future.

In addition to ORF proteins, four major structure proteins: Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N), help SARS-CoV2 in assembling and releasing new copies of the virus within human cell. We found that all high-frequent SARS-CoV2 SNVs occurred on the surface of proteins. One of most frequent mutations, D614G, has been detected to be dominant around the world now (Korber et al., 2020b). This SNV caused more infections than other mutations (Li et al., 2020b). Korber et al. (2020a,b) made suggestions from two frameworks of the potential mechanism of being more infectious: on the structure, D614G disconnects the connection between 614 in S1 and 859 in S2, which in turn facilitates the shedding of S1 from viral-membrane-bound S2 or impacts RBD-ACE2 binding by influencing RBD positioning. On the immunological aspect, D614 is within immunodominant linear epitope. Binding of antibody to the epitope may incur conformational change in Spike resulting in nearby enhanced RBD interaction with ACE2. Since furin cleavage sites are essential for SARS-CoV-2 infection of human, in addition to D614G, variants on or nearby the furin cleavage sites may affect virus entry and spread (Hoffmann et al., 2020a).

In summary, we attempted to uncover fundamental genetic patterns of SARS-CoV-2 which may help us understand functional consequences due to the viral genetic instability. Our efforts in exploring the views of SARS-CoV-2 migration and evolution in different geographical locations can be helpful to fight against the pandemic. Our findings may provide useful insights on SARS-CoV-2 replication, pathogenicity, and implications. We look forward to incorporating our results with other studies, e.g., interaction maps between SARS-CoV-2 proteins and human proteins (Gordon et al., 2020), for drug discovery, antibody design or vaccine development in near future.

DATA AVAILABILITY STATEMENT

The SARS-CoV-2 genome sequences were downloaded from GISAID, which are subject to GISAID's terms and conditions (https://www.gisaid.org/registration/terms-of-use/).

AUTHOR CONTRIBUTIONS

SL, JS, SF, KL, JL, LY, C-DH, and JW performed the research. SL, JS, and JW analyzed the data and wrote the manuscript. All authors reviewed and revised the manuscript.

FUNDING

This work was partially supported by the National Institutes of Health (Grant Number: P30CA082709) and Walther Cancer Foundation (Grant Number: 4301-80519/0187.01). Funding for open access charge: National Institutes of Health.

ACKNOWLEDGMENTS

We are grateful to scientists and researchers for depositing whole genomic sequences of Novel Pneumonia Coronavirus (SARS-CoV-2/hCoV-19/2019-nCoV) at the Global Initiative on Sharing All Influenza Data (GISAID) EpiFluTM; Thanks to GISAID database for allowing us to access the sequences for non-commercial scientific research. This manuscript has been released as a pre-print at medRxiv (Liu et al., 2020).

REFERENCES

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., and Ciccozzi, M. (2020). COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. *Med. Virol.* 92, 584–588. doi: 10.1002/jmv.25719

Becerra-Flores, M., and Cardozo, T. (2020). SARS. (-)CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int. J. Clin. Pract.* 6:e13525.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat. Struct. Biol.* 10:980. doi: 10.1038/nsb1203-980

Chen, S., Zheng, X., Zhu, J., Ding, R., Jin, Y., Zhang, W., et al. (2020). Extended ORF8 gene region is valuable in the epidemiological investigation of severe acute respiratory syndrome-similar coronavirus. *J. Infect. Dis.* 222, 223–233. doi: 10.1093/infdis/jiaa278

Cheng, V. C., Wong, S. C., Chuang, V. W., So, S. Y., Chen, J. H., Sridhar, S., et al. (2020). The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19). epidemic due to SARS-CoV-2. *J. Infect.* 81, 107–114. doi: 10.1016/j.jinf.2020.04.024

Chookajorn, T. (2020). Evolving COVID-19 conundrum and its impact. *Proc. Natl. Acad. Sci. U.S.A.* 117, 12520–12521. doi: 10.1073/pnas.2007076117

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N. G., and Decroly, E. (2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 176, 104742. doi: 10.1016/j.antiviral.2020.104742

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* 20, 533–534. doi: 10.1016/s1473-3099(20)30120-1

Eaaswarkhanth, M., Al Madhoun, A., and Al-Mulla, F. (2020). Could the D614G substitution in the SARS-CoV-2 spike (S). protein be associated with higher COVID-19 mortality? *Int. J. Infect. Dis.* 96, 459–460. doi: 10.1016/j.ijid.2020.05.071

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., et al. (2020). To mask or not to mask: modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect. Dis. Model.* 5, 293–308. doi: 10.1016/j.idm.2020.04.001

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

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020a). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9241–9243.

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020b). Reply to Sanchez-Pacheco et al., Chookajorn, and Mavian et al.: explaining phylogenetic network

analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 12524–12525. doi: 10.1073/pnas.2007433117

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 368, 779–782.

Gong, Y. N., Tsao, K. C., Hsiao, M. J., Huang, C. G., Huang, P. N., Huang, P. W., et al. (2020). SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerg. Microbes Infect.* 9, 1457–1466. doi: 10.1080/22221751.2020.1782271

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics.* 30, 2811–2812. doi: 10.1093/bioinformatics/btu393

Guan, Q., Sadykov, M., Nugmanova, R., Carr, M. J., Arold, S. T., and Pain, A. (2020). The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme. *biorxiv* [Preprint]. doi: 10.1101/2020.04.21.054221v1

Gudbjartsson, D. F., Helgason, A., Jonsson, H., Magnusson, O. T., Melsted, P., Norddahl, G. L., et al. (2020). Spread of SARS-CoV-2 in the icelandic population. *N. Engl. J. Med.* 382, 2302–2315.

Guzik, T. J., Mohiddin, S. A., Dimarco, A., Patel, V., Savvatis, K., Marelli-Berg, F. M., et al. (2020). COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options. *Cardiovasc. Res.* 116, 1666–1687.

Hassan, S. S., Choudhury, P. P., and Roy, B. (2020). SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. *Genomics* 116, 1666–1687.

Hoffmann, M., Kleine-Weber, H., and Pohlmann, S. (2020a). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell.* 78, 779.e5–784.e5.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., et al. (2020b). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 181, 271.e8–280.e8.

Huang, X., Pearce, R., and Zhang, Y. (2020). De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2. *Aging* 12, 11263–11276. doi: 10.18632/aging.103416

Issa, E., Merhi, G., Panossian, B., Salloum, T., and Tokajian, S. (2020). SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems* 5:e00266-20.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of M(pro). from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582, 289–293.

Kim, Y., Wilamowski, M., Jedrzejczak, R., Maltseva, N., Endres, M., Godzik, A., et al. (2020). *The 1.95 A Crystal Structure of the Co-factor Complex of NSP7 and the C-terminal Domain of NSP8 from SARS CoV-2*. Bethesda, MD: National Institutes of Health.

Kirkcaldy, R. D., King, B. A., and Brooks, J. T. (2020). COVID-19 and postinfection immunity: limited evidence, many remaining questions. *JAMA* [Epub ahead of print].

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020a). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *Biorxiv* [Preprint]. doi: 10.1101/2020.04.29.069054

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020b). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812.e19–827.e19.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kupferschmidt, K. (2020). *Mutations Can Reveal How the Coronavirus Moves—But They're Easy to Overinterpret*. Available online at: https://www.sciencemag.org/news/2020/03/mutations-canreveal-how-coronavirus-moves-they-re-easyoverinterpret (accessed April 15, 2020).

Kupferschmidt, K., and Cohen, J. (2020). Race to find COVID-19 treatments accelerates. *Science* 367, 1412–1413. doi: 10.1126/science.367.6485.1412

Lai, A., Bergna, A., Acciarri, C., Galli, M., and Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J. Med. Virol.* 92, 675–679. doi: 10.1002/jmv.25723

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

Li, J. Y., Liao, C. H., Wang, Q., Tan, Y. J., Luo, R., Qiu, Y., et al. (2020a). The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res.* 286:198074. doi: 10.1016/j.virusres.2020.198074

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., et al. (2020b). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* 182, 1284.e9–1294.e9.

Liu, S., Shen, J., Yang, L., Hu, C. D., and Wan, J. (2020). Distinct genetic spectrums and evolution patterns of SARS-CoV-28. *medRxiv* [Preprint].

Mavian, C., Pond, S. K., Marini, S., Magalis, B. R., Vandamme, A. M., Dellicour, S., et al. (2020). Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proc. Natl. Acad. Sci. U.S.A.* 117, 12522–12523. doi: 10.1073/pnas.2007295117

Michalska, K., Kim, Y., Jedrzejczak, R., Maltseva, N. I., Stols, L., Endres, M., et al. (2020). Crystal structures of SARS-CoV-2 ADP-ribose phosphatase (ADRP): from the apo form to ligand complexes. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.14.096081v2

Molloy, S. S., Bresnahan, P. A., Leppla, S. H., Klimpel, K. R., and Thomas, G. (1992). Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen. *J. Biol. Chem.* 267, 16396–16402.

Rambaut, A., Holmes, E., Hill, V., O'Toole, Á, McCrone, J. T., Ruis, C., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.17.046086v1

Sanchez-Pacheco, S. J., Kong, S., Pulido-Santacruz, P., Murphy, R. W., and Kubatko, L. (2020). Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. *Proc. Natl. Acad. Sci. U.S.A.* 117, 12518–12519. doi: 10.1073/pnas.2007062117

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., et al. (2020a). Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. U.S.A.* 117, 11727–11734.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., et al. (2020b). Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221–224. doi: 10.1038/s41586-020-2179-y

Shiryaev, S. A., Chernov, A. V., Golubkov, V. S., Thomsen, E. R., Chudin, E., Chee, M. S., et al. (2013). High-resolution analysis and functional mapping of cleavage sites and substrate proteins of furin in the human proteomes. *PLoS One* 8:e54290. doi: 10.1371/journal.pone.0054290

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22:30494.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* 7, 1012–1023.

Viruses CSGotICoTo (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544. doi: 10.1038/s41564-020-0695-z

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281.e6–292.e6.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., et al. (2020a). Structural, and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 181, 894.e9–904.e9.

Wang, Y., Tian, H., Zhang, L., Zhang, M., Guo, D., Wu, W., et al. (2020b). Reduction of secondary transmission of SARS-CoV-2 in households by face mask use, disinfection and social distancing: a cohort study in Beijing, China. *BMJ Glob. Health* 5:e002794. doi: 10.1136/bmjgh-2020-002794

West, R., Michie, S., Rubin, G. J., and Amlot, R. (2020). Applying principles of behaviour change to reduce SARS-CoV-2 transmission. *Nat. Hum. Behav.* 4, 451–459. doi: 10.1038/s41562-020-0887-9

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O., et al. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020a). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020b). Author Correction: a new coronavirus associated with human respiratory disease in China. *Nature* 580:E7.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* 12, 7–8. doi: 10.1038/nmeth.3213

Yuan, M., Wu, N. C., Zhu, X., Lee, C. D., So, R. T. Y., Lv, H., et al. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* 368, 630–633. doi: 10.1126/science.abb7269

Yurkovetskiy, L., Pascal, K. E., Tompkins-Tinch, C., Nyalile, T., Wang, Y., Baum, A., et al. (2020). SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. *bioRxiv* [Preprint]. doi: 10.1101/2020.07.04.187757v1

Zhang, L., Yang, J.-R., Zhang, Z., and Lin, Z. (2020a). Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission. *Medrxiv* [Preprint]. doi: 10.1101/2020.02.25.20027953v1

Zhang, C., Zheng, W., Huang, X., Bell, E. W., Zhou, X., and Zhang, Y. (2020b). Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *J. Proteome Res.* 19, 1351–1360. doi: 10.1021/acs.jproteome.0c00129

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., et al. (2020c). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.12.148726v1

Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y. P., et al. (2004). Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol. Biol.* 4:21. doi: 10.1186/1471-2148-4-21

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

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.

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.

Copyright © 2020 Liu, Shen, Fang, Li, Liu, Yang, Hu and Wan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fmicb.2020.574626

Comparative Evolutionary Patterns of *Burkholderia cenocepacia* and *B. multivorans* During Chronic Co-infection of a Cystic Fibrosis Patient Lung

*A. Amir Hassan[1,2†], Sandra C. dos Santos[2], Vaughn S. Cooper[3] and Isabel Sá-Correia[1,2]**

[1] iBB - Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, [2] Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, [3] Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

OPEN ACCESS

Edited by:
Jianying Gu,
College of Staten Island, CUNY,
United States

Reviewed by:
Joanna Goldberg,
School of Medicine, Emory University,
United States
Jonathan J. Dennis,
University of Alberta, Canada

***Correspondence:**
Professor Isabel Sá-Correia
isacorreia@tecnico.ulisboa.pt

†Present address:
A. Amir Hassan,
Laboratory of Microbiology – LM,
Department of Biochemistry and Microbiology, Ghent University, Ghent,
Belgium

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

Received: 20 June 2020
Accepted: 31 August 2020
Published: 25 September 2020

Citation:
Hassan AA, dos Santos SC, Cooper VS and Sá-Correia I (2020) Comparative Evolutionary Patterns of Burkholderia cenocepacia and B. multivorans During Chronic Co-infection of a Cystic Fibrosis Patient Lung.
Front. Microbiol. 11:574626.
doi: 10.3389/fmicb.2020.574626

During chronic respiratory infections of cystic fibrosis (CF) patients, bacteria adaptively evolve in response to the nutritional and immune environment as well as influence other infecting microbes. The present study was designed to gain insights into the genetic mechanisms underlying adaptation and diversification by the two most prevalent pathogenic species of the *Burkholderia cepacia* complex (Bcc), *B. cenocepacia* and *B. multivorans.* Herein, we study the evolution of both of these species during coinfection of a CF patient for 4.4 years using genome sequences of 9 *B. multivorans* and 11 *B. cenocepacia.* This co-infection spanned at least 3 years following initial infection by *B. multivorans* and ultimately ended in the patient's death by cepacia syndrome. Both species acquired several mutations with accumulation rates of 2.08 (*B. cenocepacia*) and 2.27 (*B. multivorans*) SNPs/year. Many of the mutated genes are associated with oxidative stress response, transition metal metabolism, defense mechanisms against antibiotics, and other metabolic alterations consistent with the idea that positive selection might be driven by the action of the host immune system, antibiotic therapy and low oxygen and iron concentrations. Two orthologous genes shared by *B. cenocepacia* and *B. multivorans* were found to be under strong selection and accumulated mutations associated with lineage diversification. One gene encodes a nucleotide sugar dehydratase involved in lipopolysaccharide O-antigen (OAg) biosynthesis (*wbil*). The other gene encodes a putative two-component regulatory sensor kinase protein required to sense and adapt to oxidative- and heavy metal-inducing stresses. This study contributes to understanding of shared and species-specific evolutionary patterns of *B. cenocepacia* and *B. multivorans* evolving in the same CF lung environment.

Keywords: *Burkholderia cepacia* complex, cystic fibrosis, chronic pulmonary infections, within-host evolution, comparative genomic analysis, *B. cenocepacia*, *B. multivorans*

INTRODUCTION

Chronic pulmonary infections are considered the leading cause for morbidity and premature death of patients at risk, as cystic fibrosis (CF) patients and immunocompromised individuals (Lipuma, 2010). Several opportunistic pathogens may populate the CF airways and the heterogeneous microbial populations usually present are difficult to eradicate (Lipuma, 2010). Among the bacterial pathogens, *Burkholderia cepacia* complex (Bcc) species are particularly feared by CF patients (Mahenthiralingam et al., 2005). Currently, the Bcc includes 24 closely related bacterial species (De Smet et al., 2015; Depoorter et al., 2016; Ong et al., 2016; Bach et al., 2017; Weber and King, 2017; Martina et al., 2018). Pulmonary infection with Bcc bacteria can involve a single species or co-infection with more than one Bcc species and, besides interpatient transmission, these bacteria can cause chronic infections and a marked decline of lung functions and decreased life expectancy. In certain cases, some Bcc species can lead to a lethal uncontrolled clinical deterioration with septicemia and necrotizing pneumonia (the "cepacia syndrome") (Mahenthiralingam et al., 2005).

During chronic infection by Bcc bacteria, the original infecting strain(s) establishes a population that evolves and diversifies into genetically and phenotypically distinct lineages. These lineages are thought to acquire multiple adaptations to the CF lung environment (Coutinho et al., 2011; Lieberman et al., 2011, 2014; Madeira et al., 2011, 2013; Mira et al., 2011; Silva et al., 2011, 2016; Moreira et al., 2014, 2017; Maldonado et al., 2016; Cabral et al., 2017; Hassan et al., 2017, 2019a; Lee et al., 2017; Nunvar et al., 2017). Over the last decade, several virulence factors and adaptive traits have been identified in Bcc species (Mahenthiralingam et al., 2005; Drevinek and Mahenthiralingam, 2010; Loutet and Valvano, 2010; Zlosnik and Speert, 2010; Madeira et al., 2011, 2013; Mira et al., 2011; Zlosnik et al., 2011, 2014; Sousa et al., 2017; Hassan et al., 2019a). Several studies have used genome-wide methods to identify genetic changes during long term infection by *B. dolosa* (Lieberman et al., 2011, 2014), *B. multivorans* (Silva et al., 2016; Diaz Caballero et al., 2018) and *B. cenocepacia* (Lee et al., 2017; Nunvar et al., 2017) within individual hosts. For example, a retrospective study involving 112 sequential *B. dolosa* isolates from 14 CF patients, identified several mutated genes involved in genetic variation within individual patients, in particular genes required for expression of surface polysaccharides, lipopolysaccharide O-antigen (OAg) biosynthesis, outer membrane components, iron scavenging, and antibiotic resistance (Lieberman et al., 2011). A subsequent study using the same sequential *B. dolosa* isolates suggested parallel bacterial evolution where different lineages may coexist for many years within a patient and identified candidate pathogenicity genes (Lieberman et al., 2014). Another comparative genomic study including 22 isolates of *B. multivorans* recovered over 20 years from CF patients also found parallel adaptive variations resulting from host related selection pressures in additional genes involved in the above referred functional categories (Silva et al., 2016). A more recent comparative genomic analysis including 32 clonal variants of *B. cenocepacia* obtained from 8 CF patients (Nunvar et al., 2017) revealed that, in addition to the aforementioned parallel mutations in gene functions previously described for *B. dolosa* and *B. multivorans* (Lieberman et al., 2011, 2014; Silva et al., 2016), genes related to transition metal metabolism are hotspots for nucleotide polymorphism (Nunvar et al., 2017). Another comparative study involving 215 genomes from serial *B. cenocepacia* isolates obtained from 16 CF patients during a 20 year-period also supports the above mentioned evolutionary trajectories during chronic Bcc infections and the well-established diverse-community model (Lee et al., 2017). This publication also reported the complete loss of chromosome III resulting in genome-size reduction as an adaptive trait of *B. cenocepacia* (Lee et al., 2017), consistent with a previously reported rare *in vivo* loss of the same mega-plasmid among *B. cenocepacia* clonal variants emerging during chronic infection (Moreira et al., 2017). Recently, a population genomic study focused on 111 *B. multivorans* isolates obtained from a Canadian CF patient showed potential parallel pathoadaptation involving genes associated with resistance toward multiple classes of antibiotics (Diaz Caballero et al., 2018).

Although the above genomic analyses have established the field of *Burkholderia* evolution during chronic infection of the CF airways, the comparison of the evolutionary patterns of strains of different species co-infecting the same CF patient is lacking. This was the objective of the present study focused on the most prevalent Bcc species among the CF community worldwide: *B. cenocepacia* and *B. multivorans* (Mahenthiralingam et al., 2005; Drevinek and Mahenthiralingam, 2010; Lipuma, 2010). Although *B. cenocepacia* has been identified as the most frequent species among the Bcc bacteria causing infections, with high potential for inter-patient transmission (Mahenthiralingam et al., 2005; Drevinek and Mahenthiralingam, 2010), in several countries *B. multivorans* has replaced *B. cenocepacia* as most frequent (Lipuma, 2010). Given that the CF environment is characterized by high pro-inflammatory cytokine levels, high antibiotic concentrations, high levels of oxidative stress and low oxygen concentration (Moriarty et al., 2007; Palmer et al., 2007; Reid et al., 2007; Williams et al., 2007), we hypothesize that the adaptive evolution of each Bcc species may vary in the same CF patient' lung-environment. This work was designed to compare the genome sequences of *B. cenocepacia* and *B. multivorans* clonal variants co-inhabiting the same host-selective environment to identify species-specific and shared evolutionary patterns. A retrospective study on twenty clonal variants derived from two ancestor strains (9 *B. multivorans* isolates and 11 *B. cenocepacia* isolates) was performed. They were sequentially retrieved from the same CF patient over a period of 4.4 years, from the onset of infection with *B. multivorans* followed by co-infection with *B. cenocepacia* until the patient's death from cepacia syndrome (Correia et al., 2008; Coutinho et al., 2011; Hassan et al., 2019a). These isolates were obtained in the major Portuguese CF treatment Center at Hospital de Santa Maria during consultation routines and pulmonary exacerbations that compelled the patient to hospitalization and intravenous therapy with gentamicin and ceftazidime (Correia et al., 2008; Coutinho et al., 2011). The design and successful execution of the present study provided relevant and useful information to contribute to a better understanding of the common and the

TABLE 1 | *Burkholderia cepacia* complex isolates examined, ordered based on the isolation date.

Bcc isolate	Isolation date	Bcc species
IST419	26 Feb 1998	*B. multivorans*
IST424	4 Jun 1998	*B. multivorans*
IST439	30 Jan 1999	*B. cenocepacia recA* lineage IIIA
IST453	19 Jul 1999	*B. multivorans*
IST455A/IST455B	1 Feb 2000	*B. multivorans*
IST461	4 Apr 2000	*B. multivorans*
IST495A/IST495B	29 May 2001	*B. multivorans*
IST4103	24 Jul 2001	*B. cenocepacia recA* lineage IIIA
IST4110	25 Sep 2001	*B. cenocepacia recA* lineage IIIA
IST4112	11 Oct 2001	*B. cenocepacia recA* lineage IIIA
IST4113	6 Nov 2001	*B. cenocepacia recA* lineage IIIA
IST4119	22 Jan 2002	*B. multivorans*
IST4116A/IST4116B	11 Feb 2002	*B. cenocepacia recA* lineage IIIA
IST4131	26 Feb 2002	*B. cenocepacia recA* lineage IIIA
IST4129	26 Mar 2002	*B. cenocepacia recA* lineage IIIA
IST4130	14 May 2002	*B. cenocepacia recA* lineage IIIA
IST4134	2 Jul 2002	*B. cenocepacia recA* lineage IIIA

They were obtained from a cystic fibrosis patient and their genomes were sequenced during this study.

specific evolutionary patterns occurring in *B. cenocepacia* and *B. multivorans* under selection in the same CF lung environment and host immune system during a co-infection that ultimately led to the cepacia syndrome.

MATERIALS AND METHODS

Bcc Bacterial Isolates and Growth Conditions

Eleven *B. cenocepacia* (*recA* lineage IIIA) sequential isolates and nine *B. multivorans* sequential isolates obtained from the same cystic fibrosis patient (patient J) (Cunha et al., 2003; Coutinho et al., 2011; Hassan et al., 2019a) were examined in this study. These isolates were collected during routine monitoring at Hospital de Santa Maria (HSM), Centro Hospitalar Lisboa Norte (CHLN) EPE, Lisbon, Portugal, from the sputum of a chronically infected CF patient who was under surveillance from February 1998 to July 2002 (Cunha et al., 2003; Correia et al., 2008; Coutinho et al., 2011; Hassan et al., 2019a; **Table 1**). These isolates were obtained from the onset of the Bcc infection until the patient's death with cepacia syndrome after 4.4 years of Bcc infection and were selected at random among the colonies isolated in selective *Burkholderia cepacia* Selectatab medium at the Hospital, at a specific date of isolation. Bacterial cultures are stored at −80°C in 1:1 (v/v) glycerol. Bacteria were grown in Lysogeny broth (LB; Conda, Pronadisa) at 37°C with shaking at 250 rpm or in LB agar plates.

Genomic DNA Sequencing, Assembly and Annotation

For genomic DNA extraction, bacterial cultures were prepared by suspending isolated colonies from LB agar plates in 3 mL LB broth, followed by overnight growth at 37°C with shaking at 250 rpm. Genomic DNA was extracted and purified using a DNeasy Blood and Tissue kit (Qiagen, Germany) according to manufacturer instructions. DNA concentration and purity were assessed using a Nanodrop ND-1000 spectrophotometer.

Genomic DNA samples of all the studied clonal variants were processed according to Illumina's instructions for generating paired-end libraries. In brief, *Burkholderia cenocepacia* IST439, IST4113, IST4129 and IST4134 were sequenced using a whole-genome shotgun sequencing strategy and Illumina Genome Analyser sequencing technology at CD Genomics (New York, NY, United States), generating short 100-bp paired-end reads with high coverage (∼300x), details of this protocol attached as **Supplementary Materials**. All of the *B. multivorans* clonal variants were sequenced on Illumina NextSeq 500 platform at the University of Pittsburgh, Pittsburgh, PA, United States; isolates of *B. cenocepacia* IST4103, IST4110, IST4112, IST4116A, IST4116B, IST4131 and IST4130 were sequenced using the 151-bp paired-end Illumina HiSeq platform at the University of New Hampshire Hubbard Center for Genomic Studies, after library preparation using a modified Illumina Nextera protocol designed for inexpensive library preparation of microbial genomes (Baym et al., 2015). Raw fastQ reads were analyzed using fastQC, which revealed that all isolates were sequenced at sufficient depth to accurately identify single nucleotide polymorphisms and indel mutations. Raw fastq paired-end files were processed for removal of Illumina adapters, trimming, and quality-based filtering using Trimmomatic v0.32 (Bolger et al., 2014).

Burkholderia multivorans IST419 and *B. cenocepacia* IST439, used as ancestor reference strains, were sequenced for a second time to generate a high-quality contiguous assembly. *B. multivorans* IST419 was likewise processed using Illumina NextSeq 500 platform producing a second sets of pair-ended reads, while *B. cenocepacia* IST439 was re-sequenced to generate a complete assembly by using a combination of single molecule, real-time (SMRT) Pacific Biosciences – PacBio reads and Illumina 100-bp paired-end reads (protocol described in **Supplementary Materials**). We used the hierarchical genome-assembly process workflow (HGAP3) to generate a completed assembly of *B. cenocepacia* IST439 and polished our assembly using the Quiver algorithm (Chin et al., 2013).

The trimmed reads of all the studied clonal variants (except *B. cenocepacia* IST439) were *de novo* assembled using VelvetOptimiser v2.2.4 (Zerbino and Birney, 2008; Zerbino, 2010) and SPAdes v3.11.1-linux (Bankevich et al., 2012) with automated optimization of the assembly parameters. The quality of the assemblies obtained was estimated by Quast v2.3 (Gurevich et al., 2013) and the best assembled outputs were automatically improved using Pilon v1.22 (Walker et al., 2014). The three trimmed data sets (two paired-end libraries and one mate-pair library) were used for scaffolding by SSPACE-standard v3.0 (Boetzer et al., 2011) followed by an automated improvement using Pilon v1.22 to gain the final draft genome sequences.

With draft genome sequences, contigs/scaffolds were reordered using MAUVE Contig Mover (Darling et al., 2010). All assembled-drafted genomes were reordered versus *B. cenocepacia* J2315 and *B. multivorans* ATCC_17616. We then annotated the completed assembly of *B. cenocepacia*

IST439 and the remaining draft genome sequences with the prokaryotic genome annotation tool prokka (v1.11) with a local, customized *Burkholderia* database of closely related Bcc strains (*B. cenocepacia* H111, *B. cenocepacia* J2315, *B. cenocepacia* ST32, *Burkholderia cenocepacia* VC1254, *B. dolosa* AU0158, *B. multivorans* ATCC_BAA-247, and *B. multivorans* ATCC 17616) (Seemann, 2014). BLAST Ring Image Generator (BRIG) was used to generate visual genome comparisons with the reference genomes of *B. cenocepacia* H111, *B. cenocepacia* J2315, *B. cenocepacia* ST32, *B. multivorans* ATCC_BAA-247, and *B. multivorans* ATCC 17616 (Alikhan et al., 2011). For enriched-functional annotation, eggNOG-mapper v4.5.1 was additionally applied for both strains (*B. cenocepacia* IST439 and *B. multivorans* IST419) that were used as references for the variant calling (Huerta-Cepas et al., 2017). This COG mapping was used to count the CDS per COG category for each species.

Variant Calling and SNP/INDEL Detections

Trimmed paired-end reads were mapped against the ancestor reference complete genome of *B. cenocepacia* IST439 (for all *B. cenocepacia* clonal variants) and/or against the ancestor reference draft genome sequence of *B. multivorans* IST419 for the corresponding clonal variants using BWA-MEM packages of Burrows-Wheeler Aligner (BWA v.0.7.10) (Li and Durbin, 2010) and NovoAlign[1]. Variants [Single nucleotide polymorphisms (SNPs) and insertion-deletion mutations (INDELs)] were called as described previously by using two independent standard variant calling pipelines; GATK and SAMtools/BCFtools toolbox (Li et al., 2009; Van der Auwera et al., 2013; Dillon et al., 2015, 2017). A brief description of this approach is included in **Supplementary Materials**.

To perform a quality control of the results obtained by the above reported method, a parallel comparison was performed using Breseq (Barrick et al., 2014), and similar predictions were found in the case of base-substitutions, with some discrepancies in indels. All putative SNPs/INDELs were then manually inspected and evaluated using the Integrative Genomics Viewer – IGV (Robinson et al., 2011; Thorvaldsdottir et al., 2013), and discarded if the BWA and Novoalign-produced alignments did not provide enough confidence (poor coverage, as described in the **Supplementary Materials**). Thus, we are confident that nearly all base-substitution and indels identified in this study were genuine events that arose during the *in vivo* evolutionary process. Finally, Functional annotation of the called variants was performed by using SnpEff v3.1 (Cingolani et al., 2012) with manual BLAST verification against the NCBI Microbes genome database.

Assessment of Population Structure and Phylogeny

In silico whole genome multi-locus sequence typing (wgMLST) (Jolley and Maiden, 2010) was performed to confirm the isolate sequence type (ST) and the clonality of all the isolates examined.

To obtain a comprehensive phylogeny of the studied clonal variants, the SNP calls and their matrices were automatically processed by using a set of algorithms previously described (Kaas et al., 2014) via the online web server CSI phylogeny v1.4[2] with default settings; min depth at SNP position – 10x, relative depth at SNP position – 10%, minimum distance between SNPs – 10 bp, min SNP quality – 30, and min read mapping quality – 25 (Kaas et al., 2014). Phylogenetic trees were then constructed using 70 and 110 variant positions among *B. cenocepacia* and *B. multivorans* clonal variants, respectively, by using FastTree (Price et al., 2010).

Seven *B. cenocepacia recA* lineage IIIA, two *B. multivorans*, and one *B. dolosa* deposited-reference genomes in *Burkholderia* genome database[3] were used to describe the phylogenetic relationship of the studied clinical isolates among the Bcc bacteria (**Supplementary Table 1**). Whole genome average nucleotide identity (ANI) was assessed by using fastANI script (Jain et al., 2018).

Genomic Features and Structural Analysis of the Ancestor *B. cenocepacia* and *B. multivorans* Strains Genomes

Burkholderia cenocepacia IST439 and *B. multivorans* IST419 were considered the ancestor reference strains and used to identify subsequent evolved mutations. Reported genomic islands (GIs: BcenGIs and GiST32-s) (Holden et al., 2009; Graindorge et al., 2012; Nunvar et al., 2017) were searched for to identify their presence in *B. cenocepacia* IST439 and *B. multivorans* IST419 genomes using basic local alignment search tool – BLAST (Altschul et al., 1990; Zhang et al., 2000).

The presence of Putative GIs was also predicted using IslandViewer 4 (Bertelli et al., 2017), unique GIs were manually confirmed and curated in the genome sequences of *B. cenocepacia* IST439 and *B. multivorans* IST419. For this confirmation, it was considered that they are unique continuous DNA regions absent in the available genome sequences of the published epidemic clones and longer than 10 kbp, flanked, at both sides, by homologous regions (especially those inserted immediately downstream of tRNA), and have lower (or higher) GC content compared with the rest of the genome (Guo et al., 2017; Nunvar et al., 2017). The absence of DNA regions homologous to these unique GIs in the genomes of epidemic clones – *B. cenocepacia* ET12 and ST32 (Holden et al., 2009; Nunvar et al., 2017) and *B. multivorans* ATCC_17616 and ATCC_BAA-247 (Komatsu et al., 2003; Johnson et al., 2015; Johnson et al., 2016) – was confirmed and visualized by Progressive MAUVE whole genome pairwise alignment and ACT/Artemis (Carver et al., 2005; Darling et al., 2010). Large DNA regions that respected those criteria but were not predicted by IslandViewer, were also manually identified and considered genuine-unique GIs.

The assembled genomes of IST419 and IST439 were also examined to identify acquired antimicrobial resistance –AMR – genes through different databases ResFinder 3.2 (Zankari, 2014)

[1]www.novocraft.com

[2]https://cge.cbs.dtu.dk/services/CSIPhylogeny/

[3]https://www.burkholderia.com/

and CARD (Jia et al., 2017) and virulence factors via VFanalyzer from VF-database (Liu et al., 2019). All the pathoadaptive associated genes obtained *in silico* were confirmed by using ABRicate version 0.5[4].

Ethics Statement

Studies involving the clinical isolates compared in this study were approved by Centro Hospitalar Lisboa Norte (CHLN)' ethics committee and the anonymity of the patient is preserved. Informed consent for the use of these isolates in research was obtained from the patient and/or the legal guardians. The patient's isolates and clinical data was collected as part of the epidemiological survey of Bcc bacteria involved in pulmonary infections among the CF patients receiving care followed at Hospital de Santa Maria (Cunha et al., 2003; Correia et al., 2008).

Data and Nucleotide Sequence Accession Numbers

DNA sequence reads for all isolates obtained by Illumina sequencing are available at the EMBL's European Nucleotide Archive (ENA) under accession number PRJEB20052 and PRJEB35836. The fully assembled and annotated genome of *B. cenocepacia* IST439 obtained by PacBio Sequencing is available at ERZ1345975, **Supplementary Table 1**. The other drafted and annotated genomes of *B. cenocepacia* and *B. multivorans* clonal variants obtained by Illumina Sequencing in the present study are available at **Supplementary Table 1**.

RESULTS

Genomic Analysis of *B. cenocepacia* and *B. multivorans* Ancestor Strains

The first isolates obtained from patient J, *B. cenocepacia* IST439 and *B. multivorans* IST419, were used as the ancestor reference strains for comparative genomic analysis. The genome of *B. cenocepacia* IST439 was sequenced using a combination of PacBio and Illumina sequences, yielding an assembly of three scaffolds corresponding to the three expected replicons (Holden et al., 2009) with a total genome size of 7.63 Mb (**Supplementary Table 2**). Plasmid pBCJ2315, present in the genome of *B. cenocepacia* J2315 strain (Holden et al., 2009), was not found in *B. cenocepacia* IST439 genome. The combination of two Illumina sequencing rounds of *B. multivorans* IST419 allowed the assembly of 76 scaffolds with a predicted genome size of ∼ 6.48 Mb (**Supplementary Table 2**) corresponding to the three expected replicons (Komatsu et al., 2003; Holden et al., 2009).

The genome sequences of the ancestor reference strains, *B. cenocepacia* IST439 and *B. multivorans* IST419, were screened for the presence of previously reported Genomic Islands (GIs), BcenGIs and GiST32-s (Holden et al., 2009; Graindorge et al., 2012; Nunvar et al., 2017). Most of these metabolic GIs were absent from both genomes but resistance GIs (in chromosome II) and pathogenic GI (Prophage BcepMu in chromosome III) were found (**Supplementary Table 3**). These shared GIs contain genes associated with arsenic resistance, antibiotic resistance, inorganic-ion and sulfate transporter, and stress response (Baldwin et al., 2004; Holden et al., 2009) and the *B. cenocepacia* low-oxygen-activated *lxa* locus (Sass et al., 2013). Another shared pathogenic GI, BcepMu prophage, contains genes involved in replication, regulation and pathogenesis (Summer et al., 2004) reported to be overexpressed in response to high doses of exogenous reactive oxygen species (ROS) (Peeters et al., 2010). Fourteen unique GIs were predicted in *B. cenocepacia* IST439 genome (MLST profile ST218) and termed BcenST218_GI1 to BcenST218_GI14 (**Supplementary Figure 1** – panel A and **Supplementary Table 4**) while 17 unique GIs were detected in *B. multivorans* IST419 (MLST profile ST836) and termed BmST836_GI1 to BmST836_GI17 (**Supplementary Figure 1** – panel B and **Supplementary Table 5**). All these unique GIs that were not shared by the *B. cenocepacia* and *B. multivorans* strains include multiple phage-related genes and/or genes associated with metabolism, replication, regulation and pathogenesis, and hypothetical membrane proteins (**Supplementary Tables 4**, **5**).

Burkholderia cenocepacia IST439 (ST-218) was found to be closely related to the German *B. cenocepacia* H111 isolate [ST1506, 99.2% average nucleotide identity (ANI)] obtained from sputum of a CF patient (Carlier et al., 2014) and more distantly to the worldwide ET-12 epidemic clones J2315 (99.0% ANI), BC7 (98.9% ANI), and K56-2 (99.0% ANI) (Holden et al., 2009; Varga et al., 2013) (**Supplementary Figure 2** and **Supplementary Table 1**). *B. multivorans* IST419 (ST-836) is more related to the Belgian clone *B. multivorans* ATCC_BAA-247 (ST650, 99.0% ANI) retrieved from sputum of a CF patient (Johnson et al., 2015; Johnson et al., 2016) than to the U.S. soil reference isolate *B. multivorans* ATCC_17616 (ST21, 96.9% ANI) (Komatsu et al., 2003) (**Supplementary Figure 2** and **Supplementary Table 1**).

Comparative Analysis of the Genome Sequences of *B. cenocepacia* and *B. multivorans* Sequential Clonal Isolates

The genomes of the eleven *B. cenocepacia* isolates and of the nine *B. multivorans* isolates were examined to identify mutations arising in the same CF patient during 4.4 years of chronic airway infection (**Figure 1** – panel A). These isolates were sequentially obtained from the onset of the Bcc infection with the *B. multivorans* followed by a co-infection with *B. cenocepacia* for at least three years until the patient's death (Correia et al., 2008; Coutinho et al., 2011; Hassan et al., 2019a). The genome of two colonies of different morphotypes (assigned as A and B) that were isolated at the same isolation date for each species [**Figure 1**, (Coutinho et al., 2011; Hassan et al., 2019a)] were also compared.

All isolates from the same species shared the same MLST type, confirming their clonal relationship (**Supplementary**

[4]https://github.com/tseemann/abricate

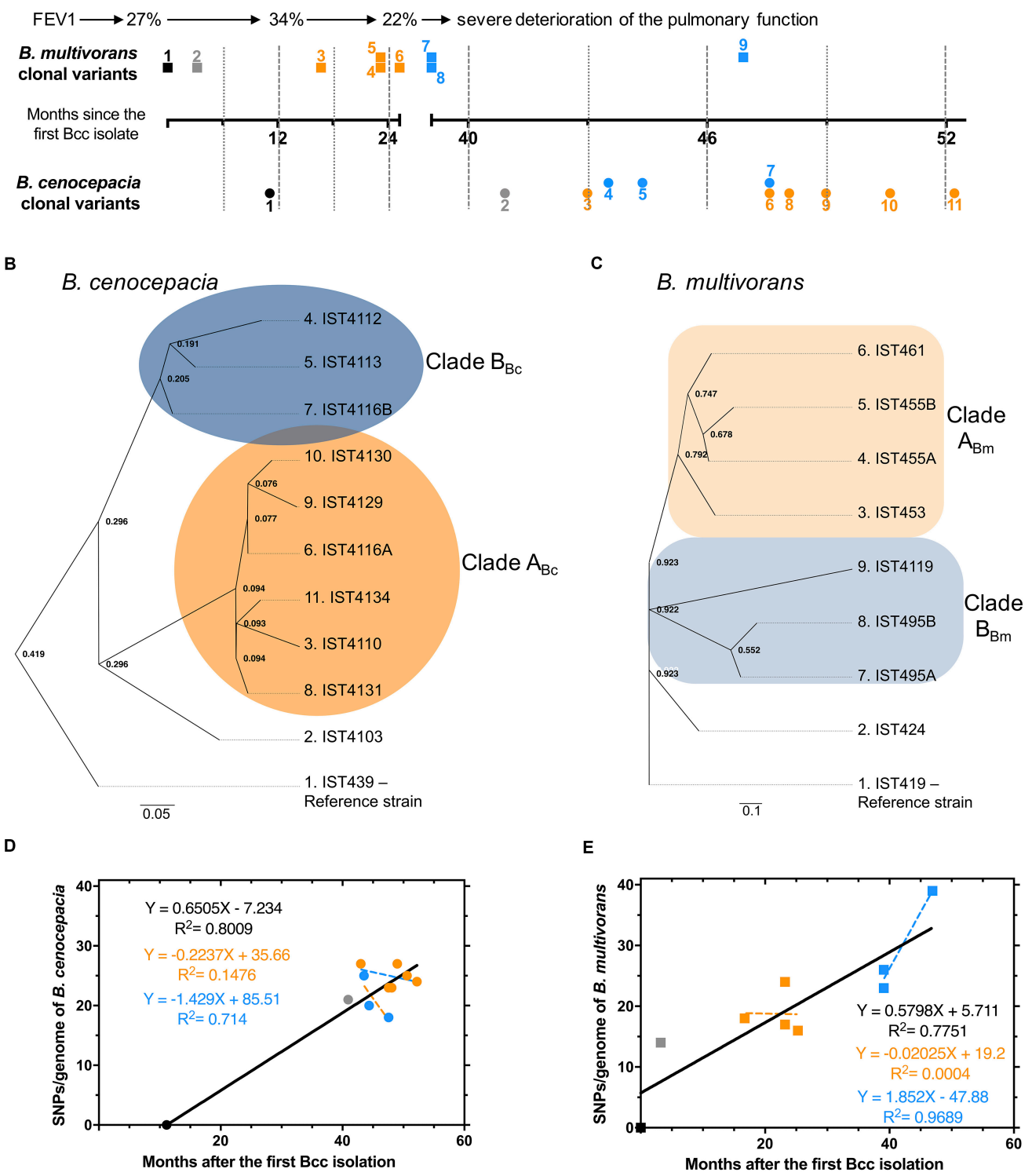


FIGURE 1 | (A) Schematic representation of the *B. cenocepacia* (•) and *B. multivorans* (■) isolates examined with information on lung function (Forced Expiratory Volume in the first second, FEV1%). Time points of isolation, in months after the first Bcc isolate was retrieved, are shown; the numerical order [also used in the phylogenetic trees of *B. cenocepacia* **(B)** and *B. multivorans* **(C)**] and the two colors used correlate with the two clades described for *B. cenocepacia* and *B. multivorans* clonal variants. **(D and E)** represent SNP accumulation during chronic infection and the relation between the time elapsed from the first Bcc isolation to the point of isolation of each isolate. A linear fit with a slope was plotted. Bc and Bm indicates *B. cenocepacia* and *B. multivorans*, respectively.

Table 2). Putative mutations were identified relative to the genomes of *B. cenocepacia* IST439 or *B. multivorans* IST419 (**Figure 1** – panels B and C), revealing 70 or 110 variant positions among *B. cenocepacia* and *B. multivorans* clonal variants, respectively. These positions were used to evaluate their phylogenetic relationships (**Figure 1**), which suggest that the two Bcc populations originated from a single colonization event followed by subsequent diversification leading to two different sub-populations (clades A_{Bc} and A_{Bm}; B_{Bc} and B_{Bm}) for each Bcc species (**Figure 1** – panels B and C – orange and blue for clade A_{Bc}, A_{Bm} and B_{Bc}, B_{Bm} – respectively). The two *B. cenocepacia* clades (clades A_{Bc} and B_{Bc}) were detected emerging during a common infection period while *B. multivorans* clade B_{Bm} was detected subsequently to clade A_{Bm} (**Figure 1** – panels A and B). The genetic differentiation among isolates did not always follow the chronology of isolation. For instance, isolates IST4110 and IST4134 are closely related although they were isolated years apart, indicating long standing coexistence of the two *B. cenocepacia* subpopulations (**Figure 1** – panels A and B). Moreover, the sole pair of *B. cenocepacia* isolates IST4116A/B obtained at the same isolation date, three years after the isolation of the first *B. cenocepacia* isolate IST439, belongs to distinct lineages (**Figure 1** – panels A and B)

Evolutionary Dynamics of *B. cenocepacia* and *B. multivorans* Co-infecting Populations Inside the CF Lung

A total of 63 single-nucleotide polymorphisms (SNPs) and 9 insertions-deletions (INDELs) were identified among *B. cenocepacia* isolates, and 97 SNPs and 16 INDELs among *B. multivorans* clonal variants (**Supplementary Tables 6, 7**). SNP accumulation rates (calculated as linear regression slope, **Figure 1** – panels D and E) indicate that the mutation rates for *B. cenocepacia* and *B. multivorans* populations were 2.08 and 2.27 SNPs/year, respectively. Most of the mutations, especially those observed along with the clades' emergence and shared among multiple isolates, are non-synonymous (**Tables 2**, **3**, **Supplementary Tables 6, 7**) and affect coding sequences (CDSs). For instance, 22 non-synonymous and 1 synonymous mutations were associated with two clades A_{Bc} and B_{Bc} that emerged within the *B. cenocepacia* population and 35 non-synonymous and 7 synonymous were related to the two clades A_{Bm} and B_{Bm} that emerged in the *B. multivorans* population. Such enrichment of non-synonymous mutations is greater than the reported substitution under neutral evolution given the codon usage and GC content of *B. cenocepacia* bacteria (Dillon et al., 2015). These results suggest that positive selection dominated during the evolution of Bcc in the CF patient lung (**Tables 2**, **3** and **Supplementary Figure 3**). Similar positive selection has been reported for other Bcc strains evolving in the CF airway, including *B. cenocepacia*, *B. multivorans* and *B. dolosa* species (Lieberman et al., 2011, 2014; Silva et al., 2016; Lee et al., 2017; Nunvar et al., 2017; Diaz Caballero et al., 2018).

Consistent with selection acting on *B. cenocepacia* clonal variants of clade B_{Bc}, only a few genes accumulated mutations that were not also found in the other clade A_{Bc} (**Table 2** and **Supplementary Table 6**). For instance, within clade B_{Bc}, the gene homologous to the IST439_04198 locus, a *penA* homolog from *B. cenocepacia* J2315 that encodes a putative beta-lactamase, was found to acquire a D262G mutation in clones IST4112, IST4113 and IST4116B. This mutation might be a result of the drastic intravenous therapy with gentamicin and ceftazidime and in agreement with the previously reported high resistance to the beta-lactam antibiotic ceftazidime (Coutinho et al., 2011). Also, among clade B_{Bc} variants, one indel was observed (P599_A601deletion) in the IST439_04244 gene, which is homologous to BCAM2211 in *B. cenocepacia* J2315 that is an ortholog of *acoR* from *Bacillus subtilis*. This *acoR* gene encodes a sigma-54 dependent transcriptional regulator that is involved in the regulation of the acetoin catabolic pathway (Huang et al., 1999; Ali et al., 2001). The acetoin catabolic pathway is believed to be a defense mechanism to avoid the lethal acidification produced by CF pathogens during fermentation and anaerobic respiration as a result of the decreased oxygen availability inside the CF lung (Whiteson et al., 2014). Moreover, other two mutations (S382L and V449L) were observed in genes homologous to IST439_00883 and IST439_00884 loci, respectively. These two loci are homologous to *orbA* and *pvdA*, respectively, that encode ornibactin siderophore biosynthesis and transport proteins (Agnoli et al., 2006). A *pvdA* deletion mutant was found to be less virulent than the parent strain in chronic and acute models of respiratory infection (Sokol et al., 1999) and a *B. cenocepacia* K56-2 *orbA* mutant (K56*orbA*:tp, *orbA* allelic exchange mutant) has also an attenuated virulence in the rat chronic respiratory infection model (Sokol et al., 2000). The reported functions of the four genes that were mutated within clade B_{Bc} suggest that positive selection might be driven by well-known environmental factors of the CF lung, in particular high antibiotic concentrations, low oxygen concentrations and low iron concentrations. On the other hand, *B. cenocepacia* clonal variants that form clade A_{Bc} accumulated mutations found in regulatory and metabolism related genes (**Table 2**). This suggests that *B. cenocepacia* tend to adapt within the CF airway by remodeling their metabolism during long term infection with periodic diversification as a function of variation within host environments. In *B. multivorans*, most of the mutated genes also affected metabolic functions (energy production and conservation and regulation of the carbohydrate metabolism) (**Table 3**).

Burkholderia cenocepacia and *B. multivorans* Genes Under Convergent Evolution in the CF Lung

The mutated genes during chronic infection belong to different clusters of orthologous groups – COG (**Supplementary Figure 3** and **Tables 2**, **3**, **Supplementary Tables 6, 7**). Genes with well-characterized functions that accumulated non-synonymous mutations (41 and 48 *B. cenocepacia* and *B. multivorans* CDSs, respectively) are schematized in **Supplementary Figure 3** and selected mutations are listed in **Tables 2**, **3**.

TABLE 2 | Summary of selected non-synonymous mutations found among *B. cenocepacia* clonal variants associated with phylogenetic clades (clades A_{Bc} and B_{Bc}, **Figure 1** – panel B, depicted in lighter and darker colors, respectively).

Gene locus	Gene prediction, Annotation	COG category	Mutation effect	Predicted functional effect and impact	presence and type of mutation in each isolate of *B. cenocepacia*									
					IST4103	IST4110	IST4112	IST4113	IST4116A	IST4116B	IST4131	IST4129	IST4130	IST4134
Cellular Processes and Signaling														
IST439_01755	wbil, UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase	Cell wall/membrane/ envelope biogenesis	L493P	CDS: missense variant, moderate			SNP	SNP		SNP				
			E487K	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_06597	*fgrL/yycG_3*, Sensor histidine kinase, Two-component regulatory system	Signal transduction mechanisms	V176M	CDS: missense variant, moderate		SNP			SNP		SNP		SNP	SNP
IST439_01475	Putative (ABC transporter) multidrug export ATP-binding membrane protein	Defense mechanisms	Q347*	CDS: stop gained, High	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_04198	*penA*, Beta-lactamase precursor		D262G	CDS: missense variant, moderate			SNP	SNP		SNP				
Information storage and processing														
IST439_00910	*bpeR/envR/tetR*, TetR family regulatory protein/HTH-type transcriptional regulator TtgR		C63R	CDS: missense variant, moderate			SNP	SNP		SNP				
			G196D	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_00921	*rbsR*, Ribose operon repressor	Transcription	V287D	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_01925	*oxyR*, Hydrogen peroxide-inducible genes activator/oxidative stress regulatory protein		S53R	CDS: missense variant, moderate			SNP	SNP		SNP				
			T669C	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_04244	*acoR_2*, Acetoin catabolism regulatory protein		P599_A601del	CDS: frameshift variant, high			INDEL	INDEL		INDEL				

(Continued)

TABLE 2 | Continued

Gene locus	Gene prediction, Annotation	COG category	Mutation effect	Predicted functional effect and impact	presence and type of mutation in each isolate of *B. cenocepacia*									
					IST4103	IST4110	IST4112	IST4113	IST4116A	IST4116B	IST4131	IST4129	IST4130	IST4134
IST439_05266	*cynR-13*, LysR family regulatory protein/HTH-type transcriptional regulator CynR		K259E	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_05755	*rpoD/sigE*, RNA polymerase sigma factor RpoD		R167T	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_06667	*yhaO/NOPBF*, putative metallophosphoesterase YhaO/putative calcineurin-like phosphoesterase family protein	Replication, recombination and repair	G153W	CDS: missense variant, moderate		SNP			SNP		SNP		SNP	SNP
Metabolism														
IST439_01911	*hmpA*, Flavohemoprotein	Energy production and conversion	D337E	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_02838	2-oxoacid ferredoxin oxidoreductase		V1062L	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_00975	*fliY_1*, Cystine-binding periplasmic protein	Amino acid transport and metabolism	D13G	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_01981	*artP/glnQ*, Arginine transporter ATP-binding subunit/Glutamine transport ATP-binding protein		Q136*	CDS: stop gained, High	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
IST439_05703	*BceL/ynfM_4,* major facilitator superfamily protein/Inner membrane transport protein YnfM	Carbohydrate transport and metabolism	T117A	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_00216	*sulP/dauA/cynT*, putative sulfate transporter/C4-dicarboxylic acid transporter DauA	Inorganic ion transport and metabolism	L61P	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_00883	*orbA*, Ornibactin receptor/Ferrichrome-iron receptor precursor		S382L	CDS: missense variant, moderate			SNP	SNP		SNP				
IST439_00884	*pvdA*, L-ornithine 5-monooxygenase	Secondary metabolites biosynthesis, transport, and catabolism	V449L	CDS: missense variant, moderate			SNP	SNP		SNP				
IST439_01356	*mdcE*, Malonate decarboxylase gamma subunit/Malonyl-S-ACP:biotin-protein carboxyltransferase MADD		D67E	CDS: missense variant, moderate		SNP			SNP		SNP	SNP	SNP	SNP
IST439_04949	*pcaC*, 4-Carboxymuconolactone decarboxylase family protein		S52R	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP

Gene loci are ordered based on COG annotation and its categories. More details in ***Supplementary Table 6****.*

TABLE 3 | Summary of selected non-synonymous mutations found among *B. multivorans* clonal variants associated with phylogenetic clades (clades A_{Bm} and B_{Bm}, **Figure 1** – panel C, depicted in lighter and darker colors, respectively).

Gene locus	Gene prediction, Annotation	COG category	Mutation effect	Predicted functional effect and impact	presence and type of mutation in each isolate of *B. multivorans*							
					IST424	IST453	IST455A	IST455B	IST461	IST495A	IST495B	IST4119
Cellular Processes and Signaling												
IST419_02260	*wbil*, UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase	Cell wall/ membrane/ envelope biogenesis	M69T (+64bp)	CDS: frameshift variant, high		INDEL	INDEL	INDEL	INDEL			
IST419_02356	*rmlB/rflB*, dTDP-glucose 4,6-dehydratase		T192H	CDS: missense variant, moderate						SNP	SNP	SNP
IST419_03735	patatin/ Putative patatin-like phospholipase		Q196*	CDS: stop gained, High						SNP	SNP	
IST419_00471	*flgD*, flagellar basal body rod modification protein	Cell motility	S219A	CDS: missense variant, moderate						SNP	SNP	
IST419_00702	*hprK*, HPr kinase/ phosphorylase	Signal transduction mechanisms	S168N	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_05375	*fgrL/ yycG_3*, Sensor histidine kinase, Two-component regulatory system		L181F							SNP	SNP	
IST419_05223	*fhaB/ palA/ hylA*, filamentous hemagglutinin outer membrane protein/ hemolysin	Intracellular trafficking, secretion, and vesicular transport	P115L	CDS: missense variant, moderate			SNP	SNP	SNP			
Information storage and processing												
IST419_01084	*rsuA/ rluF*, pseudouridine synthase, Rsu/ RNA-binding S4 domain-containing protein	Translation, ribosomal structure and biogenesis	V285A	CDS: missense variant, moderate	SNP	SNP	SNP	SNP	SNP			
Metabolism												
IST419_02576	molybdopterin oxidoreductase	Energy production and conversion	K360Q	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_05773	*tfdB*, monooxygenase FAD-binding protein		L190V	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_00066	*livF*, ABC transporter-like protein	Amino acid transport and metabolism	V185L	CDS: missense variant, moderate						SNP	SNP	

(Continued)

TABLE 3 | Continued

Gene locus	Gene prediction, Annotation	COG category	Mutation effect	Predicted functional effect and impact	presence and type of mutation in each isolate of *B. multivorans*							
					IST424	IST453	IST455A	IST455B	IST461	IST495A	IST495B	IST4119
IST419_00071	*livK*, extracellular ligand-binding receptor		Q139R	CDS: missense variant, moderate						SNP	SNP	
IST419_01537	*tcyB/ yecS*, polar amino acid ABC transporter inner membrane subunit		S165_I167dup	CDS: frameshift variant, high						INDEL	INDEL	
IST419_02455	*gltL/ aatP*, ABC transporter-like protein		D168N	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_03581	*soxA*, sarcosine oxidase subunit alpha		W385R	CDS: missense variant, moderate						SNP	SNP	
IST419_02163	*relA/ spoT*, (p)ppGpp synthetase I SpoT/ RelA	Nucleotide transport and metabolism	I493M	CDS: missense variant, moderate						SNP	SNP	
IST419_02029	*fabF2*, 3-oxoacyl-(acyl carrier protein) synthase II	Lipid transport and metabolism	G145A	CDS: missense variant, moderate						SNP	SNP	SNP
			F74T	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_05562	acetyl-CoA synthetase		A257T	CDS: missense variant, moderate	SNP		SNP	SNP				
			E484Q	CDS: missense variant, moderate		SNP			SNP	SNP	SNP	
IST419_04557	*pepM*, phosphoenolpyruvate phosphomutase	Carbohydrate transport and metabolism	D448E	CDS: missense variant, moderate						SNP	SNP	
IST419_04936	malto-oligosyltrehalose trehalohydrolase		D274N	CDS: missense variant, moderate						SNP	SNP	
IST419_03373	hypothetical methyltransferase protein	Secondary metabolites biosynthesis, transport, and catabolism	H225T	CDS: missense variant, moderate		SNP	SNP	SNP	SNP			
IST419_05081	PHB depolymerase family esterase		A272S	CDS: missense variant, moderate			SNP	SNP	SNP			

Gene loci are ordered based on COG annotation and its categories. More details in ***Supplementary Table 7****.*

The genes most likely under selection in the CF airway are those with non-synonymous mutations that become fixed within all subsequent samples. One of these is the *B. cenocepacia* gene linked to defense mechanisms against antibiotic stress (IST439_01475, encoding a putative ABC transporter ATP-binding membrane protein; homologous to BCAL1039 from *B. cenocepacia* J2315) (**Table 4**). Three other mutations affected genes involved in transcription regulation, encoding the ribose operon repressor RbsR and two sigma factor cytoplasmic proteins homologous to *B. cenocepacia* J2315 RpoC and RpoD proteins. Three mutations were also detected in genes associated with metabolism (BCAL1610 and *glnQ* homologs – encoding periplasmic cysteine-binding protein and glutamine ABC transporter ATP-binding protein – and *pcaC* homolog encoding 4-carboxymuconolactone decarboxylase, involved in the metabolism of amino acids and secondary metabolites, respectively). On the other hand, no single non-synonymous mutation became fixed among *B. multivorans* isolates during the chronic co-infection (**Tables 3** and **Supplementary Table 7**). Although the simultaneous, apparent fixation of seven new mutations over 18 months may seem high, manual inspection of the evidence supporting these mutations indicates they are genuine. It is possible that these mutations may have accumulated in a lineage evolving over a longer period of time prior to the first sampling, when clones like IST439 and more evolved lineages co-occurred.

Among the mutated genes, two genes in *B. cenocepacia* and three genes in *B. multivorans* acquired multiple independent mutations, which provide strong evidence of selection (**Figure 2**). Although multiple mutations in a single gene can result from mismapped reads and/or hidden paralogy, the evidence for these mutations is robust following manual inspection. Notably, in *B. cenocepacia*, the gene IST439_00910 that is homologous to *B. cenocepacia* J2315 *bpeR* acquired three different mutations during co-infection; one mutation (G196D) became fixed in clade A_{Bc} variants while mutation C63R emerged in clade B_{Bc} and one INDEL (K113* [-57bp]) was only found in the second sequential clonal variant isolated, IST4103, the isolate that does not belong to any of the two defined clades (**Figure 1** – panel B). The *bpeR* gene encodes a TetR family regulatory protein which is a conserved transcription regulator across the *Burkholderia* genus (Ramos et al., 2005). This gene was involved in the transcriptional control of multidrug efflux pumps, antibiotic biosynthesis, response to osmotic stress and toxic chemicals, and in pathogenicity in other Gram-negative pathogens (Ramos et al., 2005). The IST439_00910/*bpeR* is also an ortholog to *B. dolosa* AU0158 BDAG_00732 that was found before to acquire several mutations in the promotor region of 112 CF-clinical isolates (Lieberman et al., 2014). Among the three polymorphic genes in *B. multivorans* clonal variants, one polymorphic gene (IST419_01084) accumulated six different mutations and one of these variant positions (V285A) became fixed in the first 6 clonal variants (**Figure 2** – panel B). This IST419_01084 gene encodes a putative RNA pseudouridine synthase whose structure contains a small RNA-binding protein domain (S4 domain) that delivers nucleotide-modifying enzymes to RNA, regulating translation through structure specific RNA binding (Aravind and Koonin, 1999; Del Campo et al., 2001).

Orthologous Genes Experienced Positive Selection in Co-evolved *B. cenocepacia* and *B. multivorans* Populations

Two orthologous genes accumulated mutations related to lineage diversification in both *B. cenocepacia* and *B. multivorans* (**Figure 3**). These genes are homologs of *B. cenocepacia* J2315 *wbiI* and BCAS0619 and are predicted to encode a nucleotide sugar dehydratase involved in lipopolysaccharide O-antigen (OAg) biosynthesis (Ortega et al., 2005) and a putative two-component regulatory sensor kinase protein, respectively.

Regarding *wbiI*, E487K and L493P mutations were observed in IST439_01755 gene in all *B. cenocepacia* clonal variants forming clades A_{Bc} and B_{Bc}, respectively, while the M69T (+ 64bp) insertion was observed in IST419_02260 gene in all *B. multivorans* clonal variants forming clade A_{Bm}. Regardless of the Bcc species, all the isolates with a mutated *wbiI* gene expressed a lipopolysaccharide lacking the OAg (Hassan et al., 2019a). Remarkably, this *wbiI* gene was reported before to be under strong selection in *B. cenocepacia*, *B. multivorans* and *B. dolosa* during chronic lung infection (Lieberman et al., 2011, 2014; Silva et al., 2016; Hassan et al., 2017). Concerning BCAS0619, located at the third chromosome, mutation V176M was found within *B. cenocepacia* variants that form clade A_{Bc}, except for the IST4129 isolate that lacks chromosome III, while IST419_05375 gene has a L181F mutation in two *B. multivorans* variants (IST495A and IST495B).

DISCUSSION

This study was designed to gain insights into the molecular mechanisms underlying genomic diversification occurring in *B. cenocepacia* and *B. multivorans* strains that co-infected, for a period of at least 3 years, the same cystic fibrosis (CF) patient and to identify shared and relevant pathoadaptive mechanisms. Mutations with accumulation rates of 2.08 (*B. cenocepacia*) and 2.27 (*B. multivorans*) SNPs/year were found during the course of co-infection comparable with those reported for *B. dolosa* [2.1 SNPs/year] (Lieberman et al., 2011, 2014) and *B. multivorans* [2.4 SNPs/year] (Silva et al., 2016) and slightly higher than the mutation rate reported for the Czech epidemic *B. cenocepacia* ST32 strain [average rate 1.66 SNPs/year] (Nunvar et al., 2017). Such mutation rates are also comparable to those occurring in the major CF pathogen *P. aeruginosa* [2.7 SNPs/year] during CF chronic infection (Markussen et al., 2014; Marvig et al., 2015) but below the mutation rates reported for the non-Bcc member and etiological agent of melioidosis, *B. pseudomallei*, obtained from chronically infected CF patients [3.6 SNPs/year] (Viberg et al., 2017). The aforementioned mutation rates are believed to be associated with fast diversification of the infecting bacterial strains (Didelot et al., 2016) diverging into sublineages with their own functional and genomic signatures and rates of adaptation to different CF-lung environments (Markussen et al., 2014; Didelot et al., 2016).

TABLE 4 | List of non-synonymous mutations (remarked with √) that became fixed among the clonal variants of *B. cenocepacia* examined.

Gene locus/ homologous gene in J2315	Annotation	Mutational effect	COG	IST439	IST4103	IST4110	IST4112	IST4113	IST4116A	IST4116B	IST4131	IST4129	IST4130	IST4134
IST439_01475/ BCAL1039 homolog	Putative (ABC transporter) multidrug export ATP-binding/permease membrane protein	Q347*	V		√	√	√	√	√	√	√	√	√	√
IST439_00921/ *rbsR* homolog	Ribose operon repressor	V287D	K		√	√	√	√	√	√	√	√	√	√
IST439_02295/ *rpoC* homolog	DNA-directed RNA polymerase subunit beta	T669C	K		√	√	√	√	√	√	√	√	√	√
IST439_05755/ *rpoD/sigE* homolog	RpoD/SigE, RNA polymerase sigma factor RpoD	R167T	K		√	√	√	√	√	√	√	√	√	√
IST439_00975/ BCAL1610 homolog	FliY_1, Cystine-binding periplasmic protein	D13G	E		√	√	√	√	√	√	√	√	√	√
IST439_01981/ *glnQ* homolog	Glutamine transport ATP-binding protein GlnQ/ABC transporter-like protein	Q136*	E		√	√	√	√	√	√	√	√	√	√
IST439_4949/ *pcaC* homolog	PcaC, 4-Carboxymuconolactone decarboxylase family protein	S52R	Q		√	√	√	√	√	√	√	√	√	√

The only mutations presented are those acquired by CDS associated with well-characterized functions as shown in ***Supplementary Table 6****.*

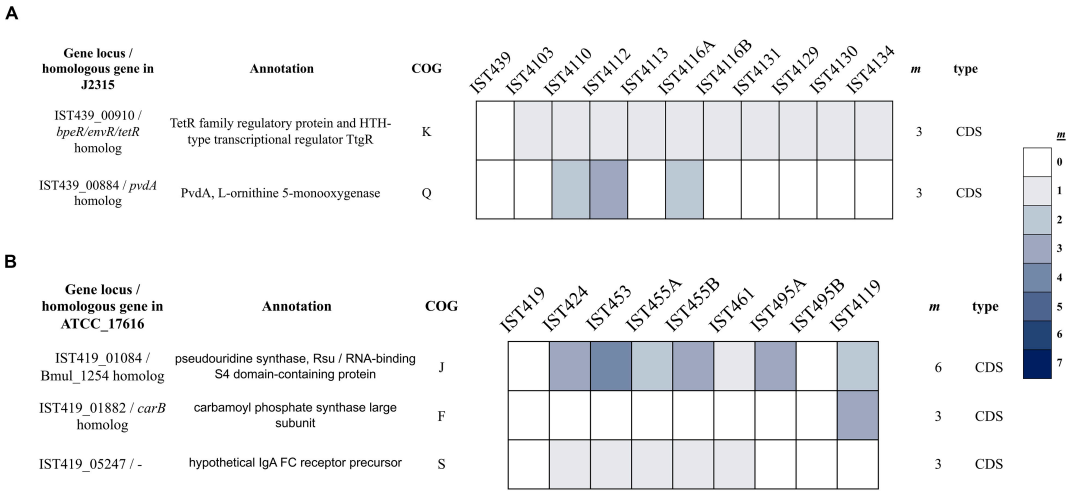


FIGURE 2 | Heat-map of non-synonymous mutation frequencies in coding sequences (CDS) under strong parallelism (that acquired ≥ 3 mutations) among the clonal variants of *B. cenocepacia* **(A)** and *B. multivorans* **(B)**. Homologous genes in *B. cenocepacia* J2315 and *B. multivorans* ATCC_17616 are indicated. Total number of mutations acquired per gene (m) and the type of mutation (coding sequence [CDS]) are presented on the right panel.

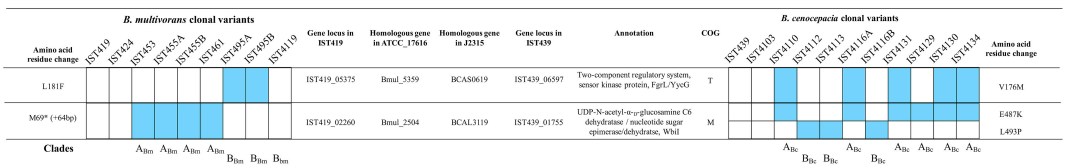


FIGURE 3 | Mutations that were observed in two orthologous genes in the genomes of clonal variants of *B. multivorans* and of *B. cenocepacia* (by chronological order from left to right) are indicated by the amino acid residue change.

In the co-infecting *B. cenocepacia* and *B. multivorans* populations examined, several genes were found to acquire different mutations either related or unrelated to lineage diversification. Remarkably, most of these genes have not been previously reported as being involved in parallel evolution in *B. dolosa*, *B. cenocepacia*, and *B. multivorans* (Lieberman et al., 2011, 2014; Silva et al., 2016; Nunvar et al., 2017). Specifically, most of the mutated genes found in this work are involved in cell envelope/wall/membrane biogenesis and in regulatory and metabolic processes, belonging to the functional categories reported before for Bcc strains as being under selective pressure in the CF lungs (Lieberman et al., 2011, 2014; Silva et al., 2016; Nunvar et al., 2017). Mutated genes with fixed non-synonymous variant positions over the entire period of chronic infection were found in the present work among the *B. cenocepacia* clonal variants examined but no single fixed non-synonymous variant position was found for *B. multivorans*. Results suggest that *B. cenocepacia* and *B. multivorans* infection involved periods of diversification dominated by positive selection, as previously suggested by others (Silva et al., 2016; Lee et al., 2017; Nunvar et al., 2017). Although we cannot discard the possibility that the spatiotemporal characteristics of the isolated clonal variants examined in this work are not representative of the population heterogeneity present at each isolation time in the CF airways, the distinct evolutionary patterns here attributed to *B. cenocepacia* and *B. multivorans* when co-infecting the same CF environment for several years, appear to be consistent and in line with the literature. Population heterogeneity in the CF lungs was observed for both species and different colony morphotypes were isolated at the same isolation dates (**Figure 1** – panels A, B and C). However, the very limited number of isolates obtained at the same isolation date and the fact that they were not always isolated randomly but to obtain different colony morphotypes limits our inferences of population dynamics. Nevertheless, heterogeneity within the *B. cenocepacia* population can be seen in the two *B cenocepacia* isolates (IST4116A/B) obtained on the same day, which belong to different lineages (clades A_{Bc} and B_{Bc}).

Among the mutated genes in *B. cenocepacia* and *B. multivorans* during long-term co-infection of the CF patient are genes encoding oxidative stress regulatory proteins and heavy metal-sensing proteins (**Tables 2**, **3** and **Supplementary Tables 6, 7**). Such regulatory proteins (e.g., OxyR and FixK) are involved in bacterial protection against reactive oxygen species (ROS) produced by phagocytic leukocytes (Reyrat et al., 1993; Saito et al., 2003; Bobik et al., 2006; Panmanee et al., 2008). Consistent with this conclusion, the host immune system was proposed to play a fundamental role in driving *B. cenocepacia* and *B. multivorans* evolution during chronic CF infection and genes linked to the oxidative stress responses (homologous to *B. cenocepacia* J2315 *katG*, *yedY*, *moeA1*, *fixL*, *osmC* and *ompR*)

reported to suffer mutation (Silva et al., 2016; Nunvar et al., 2017). This selection driven by the host immune system of genes linked to the oxidative stress and heavy metal-sensing was not however proposed for within-patient evolution of either *B. dolosa* or *P. aeruginosa* (Lieberman et al., 2011, 2014; Winstanley et al., 2016).

Increased resistance to antibiotics is also a well-described evolutionary trait associated with many CF pathogens and genes related with antibiotic resistance have been among the genes mostly reported to be under selective pressure (Jeannot et al., 2008; Lieberman et al., 2011, 2014; Silva et al., 2016; Winstanley et al., 2016; Nunvar et al., 2017). Consistent with those reports, some of the mutated genes found in the present study in *B. cenocepacia* and *B. multivorans* are involved in defense mechanisms and their regulation, as it is the case of genes encoding PenA, EmrA, ABC-2 type transporter, and DNA gyrase proteins (**Tables 2**, **3** and **Supplementary Tables 6, 7**).

Another commonly reported evolved trait during chronic infection is mutation altering genes of the OAg biosynthetic cluster (Lieberman et al., 2011, 2014; Maldonado et al., 2016; Silva et al., 2016; Hassan et al., 2017; Nunvar et al., 2017). Longitudinal comparative genomic studies of Bcc isolates from CF patients with chronic lung infection revealed the conversion from smooth (in the early isolates) to rough LPS with no OAg at the late-stages of infections by *B. cenocepacia* and *B. multivorans* (Silva et al., 2016; Hassan et al., 2017). Remarkably, an extensive study focused on *B. dolosa* during chronic lung infections has shown that late isolates produce an LPS exhibiting the OAg that was absent from the LPS of the initial infecting isolate (Lieberman et al., 2011). The phenotypic switch of OAg presence/absence was recently examined among CF isolates of several Bcc species. In the most prevalent and feared species, *B. cenocepacia* and *B. multivorans*, a marked tendency to lose the OAg along chronic infection was seen, whereas in *B. dolosa* the OAg-chain was absent from the beginning of the 5.5-year infection until the patient death (Hassan et al., 2019a). It remains important to understand why LPS is under selection in *B. cenocepacia* or *B. multivorans*. While the most common explanation invokes selection for immune evasion, it is also possible that LPS modifications are adaptations to biofilm growth (Traverse et al., 2013) or antibiotic resistance (Scribner et al., 2020). Another explanation for loss of OAg is that it increases Bcc survival in phagocytic eukaryotic cells such as amoebae, epithelial cells and human macrophages (Saldias et al., 2009; Saldias and Valvano, 2009; Kotrange et al., 2011; Maldonado et al., 2016). It is also known that the OAg absence leads to increased internalization of *B. cenocepacia* into macrophages upon phagocytosis (Saldias et al., 2009; Kotrange et al., 2011) and to facilitate *B. multivorans* growth inside macrophages (Schmerk and Valvano, 2013). Moreover, the OAg presence was found to interfere with the adhesion of *B. cenocepacia* to bronchial epithelial cells (Saldias et al., 2009) although it was reported to increase the ability of *B. cenocepacia* to adhere to atomic force microscope Si_3N_4 tip (Hassan et al., 2019b). Collectively, the loss of the OAg seems advantageous for disease development associated with *B. cenocepacia* or *B. multivorans* infections. Different mutations were accumulated in relation to the lineage emergence in the two co-infecting populations in genes IST439_01755 and IST419_02260 (homolog and ortholog to *wbiI*, respectively – **Figure 3**) that encode a nucleotide sugar dehydratase, previously found to be under parallel evolution in *B. cenocepacia* or in *B. multivorans* (Silva et al., 2016; Hassan et al., 2017). Remarkably, the comparison of the ability of three of the *B. cenocepacia* isolates examined in the present work to subvert the host's immune function, assessed by internalization assays using human dendritic cells, showed that the late variants, IST4113 and IST4134, were significantly more internalized exhibiting increased survival within dendritic cells than the early isolate IST439 (Cabral et al., 2017) the sole with an LPS with OAg (Hassan et al., 2017). Altogether, mutated genes' functions involved in the observed adaptive evolution of the two species co-inhabiting the same host-selective environment are consistent with the idea that positive selection in *B. cenocepacia* and *B. multivorans* might be driven by the action of the host immune system.

Among the few mutated regulatory proteins identified in this study is the histidine kinase protein from *B. cenocepacia* IST439_06759 (an *hisS* homolog) and from *B. multivorans* IST419_05375 that enable bacterial cells to sense and adapt to oxidative- and heavy metal- inducing stresses (Wolanin et al., 2002; Skerker et al., 2005). The orthologous genes encoding the histidine kinase protein acquired two mutations, V176M and L181F, in the two co-infecting *B. cenocepacia* and *B. multivorans* populations, respectively (**Figure 3**). This is the first time that this transcription regulator is reported to undergo adaptive evolution during CF patient infection. Signal transduction histidine kinase-related proteins are very important regulatory proteins involved in signaling cascades contributing to several simultaneous coordinated responses (Wolanin et al., 2002; Skerker et al., 2005) to, for example, osmoregulation (Tomomori et al., 1999) and chemotaxis (Bilwes et al., 1999) in bacterial pathogens.

Burkholderia cenocepacia genes encoding oxidative stress- and heavy metal-sensing- proteins were hypothesized as being the targets of convergent evolution underlying higher affinity to intra-macrophage persistence and the development of the fatal cepacia syndrome (Lee et al., 2017; Nunvar et al., 2017). Although the mechanisms underlying the fatal cepacia syndrome still remains largely unknown, the results obtained in this study are in line with those described by Nunvar et al. (2017) who hypothesized that macrophages, whose bactericidal activity relies on hydrogen peroxide and copper, are putative key players behind the development of the cepacia syndrome.

In summary, the comparative genomic analysis performed in this work identified genes which are under strong positive selection in the most prevalent and most feared Bcc species that co-infected the lung of a CF patient who died from cepacia syndrome. Among them, polymorphic genes and others involved mainly in regulatory and metabolic functions were sorted out. Since cystic fibrosis is a genetic disorder associated with inflammation, sub-optimal antioxidant protection and the continuous use of antimicrobial therapy, all resulting in marked oxidative stress (Galli et al., 2012), it is likely that in the evolved strains several physiological processes might

be activated via the global stress responses which in turn can modulate the course of infection. Global stress responses were found to be under parallel evolution in experimentally evolved *P. aeruginosa* biofilms (Winstanley et al., 2016) and were recently observed during chronic lung infections, irrespective to the presence/absence of other members of the CF microbiome (Vandeplassche et al., 2019). Given that both co-infecting Bcc species are known for their ability of intracellular survival inside macrophages (Saldias et al., 2009; Saldias and Valvano, 2009; Kotrange et al., 2011; Schmerk and Valvano, 2013), the mutated genes associated with cell envelope biogenesis, including those responsible for lipopolysaccharide OAg biosynthesis, could promote Bcc persistence through intracellular survival.

This study provides useful information to better understand the shared and species-specific evolutionary patterns of *B. cenocepacia* and *B. multivorans* evolving in the same CF lung environment during co-infection. The two orthologous genes shared by *B. cenocepacia* and *B. multivorans* that accumulate mutations associated with lineage diversification strongly suggest that two traits, lipopolysaccharide O-antigen (OAg) biosynthesis and the sensing and response to oxidative- and heavy metal-induced stresses, are biological processes under strong selection during evolution in the CF lung environment.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SS and VC carried out the first comparative analysis of *B. cenocepacia* genome sequences and AH completed this analysis and performed the genome analysis of *B. multivorans* isolates in collaboration with VC. IS-C designed and coordinated the study and wrote the manuscript with contributions from AH and VC. All authors contributed to the article and approved the submitted version.

FUNDING

Funding received by iBB—Institute for Bioengineering and Biosciences from the Portuguese Foundation for Science and Technology (FCT) (UIDB/04565/2020) is acknowledged. This work was also supported by FCT through a Ph.D. fellowship to AH (FCT-Ph.D. Programme BIOTECnico—PD/BD/113629/2015) and by NIH R01GM110444 to VSC.

ACKNOWLEDGMENTS

The authors would like to thank Carla Coutinho, Robert Sebra and Marcus Dillon for their assistance with genomic DNA extraction and PacBio sequencing and assembly.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.574626/full#supplementary-material

Supplementary Figure 1 | Genomic islands **(A)** BcenST281_GIs and **(B)** BmST836_GIs in the reference strains, *B. cenocepacia* IST439 and *B. multivorans* IST419, respectively. The inner circles denote genomes of the reference isolates and the corresponding clonal variants, ordered as indicated in the legend; Bc – *B. cenocepacia* isolates and Bm – *B. multivorans* strains. Blue and orange colors denote different clades as indicated in **Figure 1**. All clonal variants were chronologically ordered based on the isolation date. Visualizations were carried out by BRIG (Alikhan et al., 2011). For more details, see **Supplementary Tables 4, 5**.

Supplementary Figure 2 | Phylogenetic tree based on the comparison of the whole-genome of *B. cenocepacia* and *B. multivorans* clonal variants and other sequenced strains available at *Burkholderia* genome database (**Supplementary Table 1**). ANI percentages are also presented relative to each studied reference strain, *B. cenocepacia* IST439 and *B. multivorans* IST419.

Supplementary Figure S3 | 3D plot of non-synonymous mutations observed in *B. cenocepacia* **(A)** and *B. multivorans* **(B)** clonal variants during the chronic co-infection. All mutated genes of well-characterized functions are depicted as green oval shapes (with SNPs) and/or blue shapes (with INDELs). The number of mutated genes per each COG category is illustrated above each corresponding shape as a ratio (number of genes per isolate/number of all mutated genes in the entire population for this particular COG). Accumulation of small green shapes indicates the number of SNP mutations that was observed per isolate. For a detailed list of mutations, see **Supplementary Tables 6, 7** for *B. cenocepacia* and *B. multivorans*, respectively. Abbreviation letters representing the COGs are detailed in the bottom panel.

Supplementary Table S1 | List of the whole genome sequences of Bcc isolates used in the present study for comparative genomic analyses. The isolation year and country are indicated as well as the multi-locus sequence typing (MLST). CS – cepacia syndrome.

Supplementary Table S2 | Metadata information of the studied clonal variants sequentially ordered based on the isolation date; details on genomic assembly and functional annotation are shown as well as the *in silico* MLST profiles.

Supplementary Table S3 | Published genomic islands (Holden et al., 2009; Graindorge et al., 2012; Nunvar et al., 2017) present in *B. cenocepacia* J2315 and ST32 and their eventual presence in *B. cenocepacia* IST439 and *B. multivorans* IST419.

Supplementary Table S4 | BcenST218_GIs as Genomic islands of *B. cenocepacia* IST439 and their presence in *B. multivorans* IST419, *B. cenocepacia* J2315 and ST32, *B. multivorans* ATCC_17616 and ATCC_BAA-247.

Supplementary Table S5 | BmST836_GIs as Genomic islands of *B. multivorans* IST419 and their presence in *B. cenocepacia* IST439, *B. cenocepacia* J2315 and ST32, *B. multivorans* ATCC_17616 and ATCC_BAA-247.

Supplementary Table S6 | List of mutations found among *B. cenocepacia* clonal variants. All mutated genes were grouped to predicted gene functions (COG). Isolates are chronologically ordered as indicated in **Figure 1**. A – GATK pipeline and B – SAMtools/BCFtools toolbox.

Supplementary Table S7 | List of mutations found among *B. multivorans* clonal variants. All mutated genes were grouped to predicted gene functions (COG). Isolates are chronologically ordered as indicated in **Figure 1**. A – GATK pipeline and B – SAMtools/BCFtools toolbox.

REFERENCES

Agnoli, K., Lowe, C. A., Farmer, K. L., Husnain, S. I., and Thomas, M. S. (2006). The ornibactin biosynthesis and transport genes of *Burkholderia cenocepacia* are regulated by an extracytoplasmic function sigma factor which is a part of the Fur regulon. *J. Bacteriol.* 188, 3631–3644. doi: 10.1128/jb.188.10.3631-3644.2006

Ali, N. O., Bignon, J., Rapoport, G., and Debarbouille, M. (2001). Regulation of the acetoin catabolic pathway is controlled by sigma L in *Bacillus subtilis*. *J. Bacteriol.* 183, 2497–2504. doi: 10.1128/JB.183.8.2497-2504.2001

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Aravind, L., and Koonin, E. V. (1999). Novel predicted RNA-binding domains associated with the translation machinery. *J. Mol. Evol.* 48, 291–302.

Bach, E., Sant'Anna, F. H., Magrich Dos Passos, J. F., Balsanelli, E., de Baura, V. A., Pedrosa, F. O., et al. (2017). Detection of misidentifications of species from the *Burkholderia cepacia* complex and description of a new member, the soil bacterium *Burkholderia catarinensis* sp. nov. *Pathogens Dis.* 75:ftx076. doi: 10.1093/femspd/ftx076

Baldwin, A., Sokol, P. A., Parkhill, J., and Mahenthiralingam, E. (2004). The *Burkholderia cepacia* epidemic strain marker is part of a novel genomic island encoding both virulence and metabolism-associated genes in *Burkholderia cenocepacia*. *Infect. Immun.* 72, 1537–1547. doi: 10.1128/iai.72.3.1537-1547.2004

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. Comp. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., et al. (2014). Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics* 15:1039. doi: 10.1186/1471-2164-15-1039

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., and Kishony, R. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* 10:e0128036. doi: 10.1371/journal.pone.0128036

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research, Computing, Group, Lau, B. Y., et al. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45, W30–W35. doi: 10.1093/nar/gkx343

Bilwes, A. M., Alex, L. A., Crane, B. R., and Simon, M. I. (1999). Structure of CheA, a signal-transducing histidine kinase. *Cell* 96, 131–141. doi: 10.1016/s0092-8674(00)80966-6

Bobik, C., Meilhoc, E., and Batut, J. (2006). FixJ: a major regulator of the oxygen limitation response and late symbiotic functions of *Sinorhizobium meliloti*. *J. Bacteriol.* 188, 4890–4902. doi: 10.1128/JB.00251-06

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

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

Cabral, M., Pereira, M., Silva, Z., Iria, I., Coutinho, C., Lopes, A., et al. (2017). Using dendritic cells to evaluate how *Burkholderia cenocepacia* clonal isolates from a chronically infected cystic fibrosis patient subvert immune functions. *Med. Microbiol. Immunol.* 206, 111–123. doi: 10.1007/s00430-016-0488-4

Carlier, A., Agnoli, K., Pessi, G., Suppiger, A., Jenul, C., Schmid, N., et al. (2014). Genome sequence of *Burkholderia cenocepacia* H111, a cystic fibrosis airway isolate. *Genome Announc.* 2:e00298-14. doi: 10.1128/genomeA.00298-14

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., and Parkhill, J. (2005). ACT: the artemis comparison tool. *Bioinformatics* 21, 3422–3423. doi: 10.1093/bioinformatics/bti553

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10, 563–569. doi: 10.1038/nmeth.2474

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

Correia, S., Nascimento, C., Pereira, L., Cunha, M. V., Sá-Correia, I., and Barreto, C. (2008). Infecção respiratória por bactérias do complexo *cepacia*: evolução clínica em doentes com fibrose quística. *Rev. Portuguesa Pneumol.* 14, 5–26.

Coutinho, C. P., de Carvalho, C. C., Madeira, A., Pinto-de-Oliveira, A., and Sa-Correia, I. (2011). *Burkholderia cenocepacia* phenotypic clonal variation during a 3.5-year colonization in the lungs of a cystic fibrosis patient. *Infect. Immun.* 79, 2950–2960. doi: 10.1128/IAI.01366-10

Cunha, M. V., Leitão, J. H., Mahenthiralingam, E., Vandamme, P., Lito, L., Barreto, C., et al. (2003). Molecular analysis of *Burkholderia cepacia* complex isolates from a Portuguese cystic fibrosis center: a 7-year study. *J. Clin. Microbiol.* 41, 4113–4120. doi: 10.1128/JCM.41.9.4113-4120.2003

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

De Smet, B., Mayo, M., Peeters, C., Zlosnik, J. E., Spilker, T., Hird, T. J., et al. (2015). *Burkholderia stagnalis* sp. nov. and *Burkholderia territorii* sp. nov., two novel *Burkholderia cepacia* complex species from environmental and human sources. *Int. J. Syst. Evol. Microbiol.* 65, 2265–2271. doi: 10.1099/ijs.0.000251

Del Campo, M., Kaya, Y., and Ofengand, J. (2001). Identification and site of action of the remaining four putative pseudouridine synthases in *Escherichia coli*. *Rna* 7, 1603–1615.

Depoorter, E., Bull, M. J., Peeters, C., Coenye, T., Vandamme, P., and Mahenthiralingam, E. (2016). *Burkholderia*: an update on taxonomy and biotechnological potential as antibiotic producers. *Appl. Microbiol. Biotechnol.* 100, 5215–5229. doi: 10.1007/s00253-016-7520-x

Diaz Caballero, J., Clark, S. T., Wang, P. W., Donaldson, S. L., Coburn, B., Tullis, D. E., et al. (2018). A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in *Burkholderia multivorans*. *PLoS Pathog.* 14:e1007453. doi: 10.1371/journal.ppat.1007453

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

Dillon, M. M., Sung, W., Lynch, M., and Cooper, V. S. (2015). The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. *Genetics* 200, 935–946. doi: 10.1534/genetics.115.176834

Dillon, M. M., Sung, W., Sebra, R., Lynch, M., and Cooper, V. S. (2017). Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol. Biol. Evol.* 34, 93–109. doi: 10.1093/molbev/msw224

Drevinek, P., and Mahenthiralingam, E. (2010). *Burkholderia cenocepacia* in cystic fibrosis: epidemiology and molecular mechanisms of virulence. *Clin. Microbiol. Infect.* 16, 821–830. doi: 10.1111/j.1469-0691.2010.03237.x

Galli, F., Battistoni, A., Gambari, R., Pompella, A., Bragonzi, A., Pilolli, F., et al. (2012). Oxidative stress and antioxidant therapy in cystic fibrosis. *Biochim. Biophys. Acta (BBA)* 1822, 690–713. doi: 10.1016/j.bbadis.2011.12.012

Graindorge, A., Menard, A., Monnez, C., and Cournoyer, B. (2012). Insertion sequence evolutionary patterns highlight convergent genetic inactivations and recent genomic island acquisitions among epidemic *Burkholderia cenocepacia*. *J. Med. Microbiol.* 61(Pt 3), 394–409. doi: 10.1099/jmm.0.036822-0

Guo, F. B., Xiong, L., Zhang, K. Y., Dong, C., Zhang, F. Z., and Woo, P. C. (2017). Identification and analysis of genomic islands in *Burkholderia cenocepacia* AU 1054 with emphasis on pathogenicity islands. *BMC Microbiol.* 17:73. doi: 10.1186/s12866-017-0986-6

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

Hassan, A. A., Coutinho, C. P., and Sá-Correia, I. (2019a). *Burkholderia cepacia* complex species differ in the frequency of variation of the lipopolysaccharide O-antigen expression during cystic fibrosis chronic respiratory infection. *Front. Cell. Infect. Microbiol.* 9:273. doi: 10.3389/fcimb.2019.00273

Hassan, A. A., Maldonado, R. F., Dos Santos, S. C., Di Lorenzo, F., Silipo, A., Coutinho, C. P., et al. (2017). Structure of O-antigen and hybrid biosynthetic locus in *Burkholderia cenocepacia* clonal variants recovered from a cystic fibrosis patient. *Front. Microbiol.* 8:1027. doi: 10.3389/fmicb.2017.01027

Hassan, A. A., Vitorino, M. V., Robalo, T., Rodrigues, M. S., and Sa-Correia, I. (2019b). Variation of *Burkholderia cenocepacia* cell wall morphology and mechanical properties during cystic fibrosis lung infection, assessed by atomic force microscopy. *Sci. Rep.* 9:16118. doi: 10.1038/s41598-019-52604-9

Holden, M. T., Seth-Smith, H. M., Crossman, L. C., Sebaihia, M., Bentley, S. D., Cerdeno-Tarraga, A. M., et al. (2009). The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J. Bacteriol.* 191, 261–277. doi: 10.1128/JB.01230-08

Huang, M., Oppermann-Sanio, F. B., and Steinbuchel, A. (1999). Biochemical and molecular characterization of the *Bacillus subtilis* acetoin catabolic pathway. *J. Bacteriol.* 181, 3837–3841.

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

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

Jeannot, K., Elsen, S., Kohler, T., Attree, I., van Delden, C., and Plesiat, P. (2008). Resistance and virulence of *Pseudomonas aeruginosa* clinical strains overproducing the MexCD-OprJ efflux pump. *Antimicrob. Agents Chemother.* 52, 2455–2462. doi: 10.1128/AAC.01107-07

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Johnson, S. L., Bishop-Lilly, K. A., Ladner, J. T., Daligault, H. E., Davenport, K. W., Jaissle, J., et al. (2016). Correction for Johnson et al., complete genome sequences for 59 *Burkholderia* isolates, both pathogenic and near neighbor. *Genome Announc.* 4:e00313-16. doi: 10.1128/genomeA.00313-16

Johnson, S. L., Bishop-Lilly, K. A., Ladner, J. T., Daligault, H. E., Davenport, K. W., Jaissle, J., et al. (2015). Complete genome sequences for 59 *burkholderia* isolates, both pathogenic and near neighbor. *Genome Announc.* 3:e00159-15. doi: 10.1128/genomeA.00159-15

Jolley, K. A., and Maiden, M. C. J. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595. doi: 10.1186/1471-2105-11-595

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., and Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One* 9:e104984. doi: 10.1371/journal.pone.0104984

Komatsu, H., Imura, Y., Ohori, A., Nagata, Y., and Tsuda, M. (2003). Distribution and organization of auxotrophic genes on the multichromosomal genome of *Burkholderia multivorans* ATCC 17616. *J. Bacteriol.* 185, 3333–3343. doi: 10.1128/JB.185.11.3333-3343.2003

Kotrange, S., Kopp, B., Akhter, A., Abdelaziz, D., Abu Khweek, A., Caution, K., et al. (2011). *Burkholderia cenocepacia* O polysaccharide chain contributes to caspase-1-dependent IL-1beta production in macrophages. *J. Leukoc. Biol.* 89, 481–488. doi: 10.1189/jlb.0910513

Lee, A. H., Flibotte, S., Sinha, S., Paiero, A., Ehrlich, R. L., Balashov, S., et al. (2017). Phenotypic diversity and genotypic flexibility of *Burkholderia cenocepacia* during long-term chronic infection of cystic fibrosis lungs. *Genome Res.* 27, 650–662. doi: 10.1101/gr.213363.116

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

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

Lieberman, T. D., Flett, K. B., Yelin, I., Martin, T. R., McAdam, A. J., Priebe, G. P., et al. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature Genet.* 46, 82–87. doi: 10.1038/ng.2848

Lieberman, T. D., Michel, J. B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M. R. Jr., et al. (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genet.* 43, 1275–1280. doi: 10.1038/ng.997

Lipuma, J. J. (2010). The changing microbial epidemiology in cystic fibrosis. *Clin. Microbiol. Rev.* 23, 299–323. doi: 10.1128/CMR.00068-09

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

Loutet, S. A., and Valvano, M. A. (2010). A decade of *Burkholderia cenocepacia* virulence determinant research. *Infect. Immun.* 78, 4088–4100. doi: 10.1128/IAI.00212-10

Madeira, A., dos Santos, S. C., Santos, P. M., Coutinho, C. P., Tyrrell, J., McClean, S., et al. (2013). Proteomic profiling of *Burkholderia cenocepacia* clonal isolates with different virulence potential retrieved from a cystic fibrosis patient during chronic lung infection. *PLoS One* 8:e83065. doi: 10.1371/journal.pone.0083065

Madeira, A., Santos, P. M., Coutinho, C. P., Pinto-de-Oliveira, A., and Sá-Correia, I. (2011). Quantitative proteomics (2-D DIGE) reveals molecular strategies employed by *Burkholderia cenocepacia* to adapt to the airways of cystic fibrosis patients under antimicrobial therapy. *Proteomics* 11, 1313–1328. doi: 10.1002/pmic.201000457

Mahenthiralingam, E., Urban, T. A., and Goldberg, J. B. (2005). The multifarious, multireplicon *Burkholderia cepacia* complex. *Nature Rev. Microbiol.* 3, 144–156. doi: 10.1038/nrmicro1085

Maldonado, R. F., Sá-Correia, I., and Valvano, M. A. (2016). Lipopolysaccharide modification in Gram-negative bacteria during chronic infection. *FEMS Microbiol. Rev.* 40, 480–493. doi: 10.1093/femsre/fuw007

Markussen, T., Marvig, R. L., Gomez-Lozano, M., Aanaes, K., Burleigh, A. E., Hoiby, N., et al. (2014). Environmental heterogeneity drives within-host diversification and evolution of *Pseudomonas aeruginosa*. *mBio* 5:e01592-14. doi: 10.1128/mBio.01592-14

Martina, P., Leguizamon, M., Prieto, C. I., Sousa, S. A., Montanaro, P., Draghi, W. O., et al. (2018). *Burkholderia puraquae* sp. nov., a novel species of the *Burkholderia cepacia* complex isolated from hospital settings and agricultural soils. *Int. J. Syst. Evol. Microbiol.* 68, 14–20. doi: 10.1099/ijsem.0.002293

Marvig, R. L., Dolce, D., Sommer, L. M., Petersen, B., Ciofu, O., Campana, S., et al. (2015). Within-host microevolution of *Pseudomonas aeruginosa* in Italian cystic fibrosis patients. *BMC Microbiol.* 15:218. doi: 10.1186/s12866-015-0563-9

Mira, N. P., Madeira, A., Moreira, A. S., Coutinho, C. P., and Sá-Correia, I. (2011). Genomic expression analysis reveals strategies of *Burkholderia cenocepacia* to adapt to cystic fibrosis patients' airways and antimicrobial therapy. *PLoS One* 6:e28831. doi: 10.1371/journal.pone.0028831

Moreira, A. S., Coutinho, C. P., Azevedo, P., Lito, L., Melo-Cristino, J., and Sá-Correia, I. (2014). *Burkholderia dolosa* phenotypic variation during the decline in lung function of a cystic fibrosis patient during 5.5 years of chronic colonization. *J. Med. Microbiol.* 63(Pt 4), 594–601. doi: 10.1099/jmm.0.069849-0

Moreira, A. S., Mil-Homens, D., Sousa, S. A., Coutinho, C. P., Pinto-de-Oliveira, A., Ramos, C. G., et al. (2017). Variation of *Burkholderia cenocepacia* virulence potential during cystic fibrosis chronic lung infection. *Virulence* 8, 782–796. doi: 10.1080/21505594.2016.1237334

Moriarty, T. F., McElnay, J. C., Elborn, J. S., and Tunney, M. M. (2007). Sputum antibiotic concentrations: implications for treatment of cystic fibrosis lung infection. *Pediatric Pulmonol.* 42, 1008–1017. doi: 10.1002/ppul.20671

Nunvar, J., Capek, V., Fiser, K., Fila, L., and Drevinek, P. (2017). What matters in chronic *Burkholderia cenocepacia* infection in cystic fibrosis: insights from comparative genomics. *PLoS Pathog.* 13:e1006762. doi: 10.1371/journal.ppat.1006762

Ong, K. S., Aw, Y. K., Lee, L. H., Yule, C. M., Cheow, Y. L., and Lee, S. M. (2016). *Burkholderia paludis* sp nov., an antibiotic-siderophore producing novel *Burkholderia cepacia* complex species, isolated from Malaysian tropical peat swamp soil. *Front. Microbiol.* 7:2046. doi: 10.3389/fmicb.2016.02046

Ortega, X., Hunt, T. A., Loutet, S., Vinion-Dubiel, A. D., Datta, A., Choudhury, B., et al. (2005). Reconstitution of O-specific lipopolysaccharide expression in *Burkholderia cenocepacia* strain J2315, which is associated with transmissible infections in patients with cystic fibrosis. *J. Bacteriol.* 187, 1324–1333. doi: 10.1128/JB.187.4.1324-1333.2005

Palmer, K. L., Aye, L. M., and Whiteley, M. (2007). Nutritional cues control *Pseudomonas aeruginosa* multicellular behavior in cystic fibrosis sputum. *J. Bacteriol.* 189, 8079–8087. doi: 10.1128/JB.01138-07

Panmanee, W., Gomez, F., Witte, D., Pancholi, V., Britigan, B. E., and Hassett, D. J. (2008). The peptidoglycan-associated lipoprotein OprL helps protect a *Pseudomonas aeruginosa* mutant devoid of the transactivator OxyR from

hydrogen peroxide-mediated killing during planktonic and biofilm culture. *J. Bacteriol.* 190, 3658–3669. doi: 10.1128/JB.00022-08

Peeters, E., Sass, A., Mahenthiralingam, E., Nelis, H., and Coenye, T. (2010). Transcriptional response of *Burkholderia cenocepacia* J2315 sessile cells to treatments with high doses of hydrogen peroxide and sodium hypochlorite. *BMC Genomics* 11:90. doi: 10.1186/1471-2164-11-90

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., et al. (2005). The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev. MMBR* 69, 326–356. doi: 10.1128/MMBR.69.2.326-356.2005

Reid, D. W., Misso, N., Aggarwal, S., Thompson, P. J., and Walters, E. H. (2007). Oxidative stress and lipid-derived inflammatory mediators during acute exacerbations of cystic fibrosis. *Respirology* 12, 63–69. doi: 10.1111/j.1440-1843.2006.00962.x

Reyrat, J. M., David, M., Blonski, C., Boistard, P., and Batut, J. (1993). Oxygen-regulated in vitro transcription of *Rhizobium meliloti* nifA and fixK genes. *J. Bacteriol.* 175, 6867–6872. doi: 10.1128/jb.175.21.6867-6872.1993

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

Saito, K., Ito, E., Hosono, K., Nakamura, K., Imai, K., Iizuka, T., et al. (2003). The uncoupling of oxygen sensing, phosphorylation signalling and transcriptional activation in oxygen sensor FixL and FixJ mutants. *Mol. Microbiol.* 48, 373–383.

Saldias, M. S., Ortega, X., and Valvano, M. A. (2009). *Burkholderia cenocepacia* O antigen lipopolysaccharide prevents phagocytosis by macrophages and adhesion to epithelial cells. *J. Med. Microbiol.* 58(Pt 12), 1542–1548. doi: 10.1099/jmm.0.013235-0

Saldias, M. S., and Valvano, M. A. (2009). Interactions of *Burkholderia cenocepacia* and other *Burkholderia cepacia* complex bacteria with epithelial and phagocytic cells. *Microbiology* 155(Pt 9), 2809–2817. doi: 10.1099/mic.0.031344-0

Sass, A. M., Schmerk, C., Agnoli, K., Norville, P. J., Eberl, L., Valvano, M. A., et al. (2013). The unexpected discovery of a novel low-oxygen-activated locus for the anoxic persistence of *Burkholderia cenocepacia*. *ISME J.* 7, 1568–1581. doi: 10.1038/ismej.2013.36

Schmerk, C. L., and Valvano, M. A. (2013). *Burkholderia multivorans* survival and trafficking within macrophages. *J. Med. Microbiol.* 62(Pt 2), 173–184. doi: 10.1099/jmm.0.051243-0

Scribner, M. R., Santos-Lopez, A., Marshall, C. W., Deitrick, C., and Cooper, V. S. (2020). Parallel evolution of tobramycin resistance across species and environments. *mBio* 11:e00932-20. doi: 10.1128/mBio.00932-20

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

Silva, I. N., Ferreira, A. S., Becker, J. D., Zlosnik, J. E., Speert, D. P., He, J., et al. (2011). Mucoid morphotype variation of *Burkholderia multivorans* during chronic cystic fibrosis lung infection is correlated with changes in metabolism, motility, biofilm formation and virulence. *Microbiology* 157(Pt 11), 3124–3137. doi: 10.1099/mic.0.050989-0

Silva, I. N., Santos, P. M., Santos, M. R., Zlosnik, J. E., Speert, D. P., Buskirk, S. W., et al. (2016). Long-term evolution of *Burkholderia multivorans* during a chronic cystic fibrosis infection reveals shifting forces of selection. *mSystems* 1:e00029-16. doi: 10.1128/mSystems.00029-16

Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G., and Laub, M. T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol.* 3:e334. doi: 10.1371/journal.pbio.0030334

Sokol, P. A., Darling, P., Lewenza, S., Corbett, C. R., and Kooi, C. D. (2000). Identification of a siderophore receptor required for ferric ornibactin uptake in *Burkholderia cepacia*. *Infect. Immun.* 68, 6554–6560. doi: 10.1128/iai.68.12.6554-6560.2000

Sokol, P. A., Darling, P., Woods, D. E., Mahenthiralingam, E., and Kooi, C. (1999). Role of ornibactin biosynthesis in the virulence of *Burkholderia cepacia*: characterization of pvdA, the gene encoding L-ornithine N(5)-oxygenase. *Infect. Immun.* 67, 4443–4455.

Sousa, S. A., Feliciano, J. R., Pita, T., Guerreiro, S. I., and Leitao, J. H. (2017). *Burkholderia cepacia* complex regulation of virulence gene expression: a review. *Genes (Basel)* 8:43. doi: 10.3390/genes8010043

Summer, E. J., Gonzalez, C. F., Carlisle, T., Mebane, L. M., Cass, A. M., Savva, C. G., et al. (2004). *Burkholderia cenocepacia* phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. *J. Mol. Biol.* 340, 49–65. doi: 10.1016/j.jmb.2004.04.053

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

Tomomori, C., Tanaka, T., Dutta, R., Park, H., Saha, S. K., Zhu, Y., et al. (1999). Solution structure of the homodimeric core domain of *Escherichia coli* histidine kinase EnvZ. *Nature Struct. Mol. Biol.* 6, 729–734. doi: 10.1038/11495

Traverse, C. C., Mayo-Smith, L. M., Poltak, S. R., and Cooper, V. S. (2013). Tangled bank of experimentally evolved *Burkholderia* biofilms reflects selection during chronic infections. *Proc. Natl. Acad. Sci. U.S.A.* 110, E250–E259. doi: 10.1073/pnas.1207025110

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Vandeplassche, E., Sass, A., Lemarcq, A., Dandekar, A. A., Coenye, T., and Crabbe, A. (2019). In vitro evolution of *Pseudomonas aeruginosa* AA2 biofilms in the presence of cystic fibrosis lung microbiome members. *Sci. Rep.* 9:12859. doi: 10.1038/s41598-019-49371-y

Varga, J. J., Losada, L., Zelazny, A. M., Kim, M., McCorrison, J., Brinkac, L., et al. (2013). Draft genome sequences of *Burkholderia cenocepacia* ET12 lineage strains K56-2 and BC7. *Genome Announc.* 1:e00841-13. doi: 10.1128/genomeA.00841-13

Viberg, L. T., Sarovich, D. S., Kidd, T. J., Geake, J. B., Bell, S. C., Currie, B. J., et al. (2017). Within-host evolution of *Burkholderia pseudomallei* during chronic infection of seven Australasian cystic fibrosis patients. *mBio* 8:e00356-17. doi: 10.1128/mBio.00356-17

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

Weber, C. F., and King, G. M. (2017). Volcanic Soils as Sources of Novel CO-Oxidizing *Paraburkholderia* and *Burkholderia*: *Paraburkholderia hiiakae* sp. nov., *Paraburkholderia metrosideri* sp. nov., *Paraburkholderia paradisi* sp. nov., *Paraburkholderia peleae* sp. nov., and *Burkholderia alpina* sp. nov. a member of the *Burkholderia cepacia* complex. *Front. Microbiol.* 8:207. doi: 10.3389/fmicb.2017.00207

Whiteson, K. L., Meinardi, S., Lim, Y. W., Schmieder, R., Maughan, H., Quinn, R., et al. (2014). Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. *ISME J.* 8, 1247–1258. doi: 10.1038/ismej.2013.229

Williams, H. D., Zlosnik, J. E., and Ryall, B. (2007). Oxygen, cyanide and energy generation in the cystic fibrosis pathogen *Pseudomonas aeruginosa*. *Adv. Microb. Physiol.* 52, 1–71. doi: 10.1016/S0065-2911(06)52001-6

Winstanley, C., O'Brien, S., and Brockhurst, M. A. (2016). *Pseudomonas aeruginosa* evolutionary adaptation and diversification in cystic fibrosis chronic lung infections. *Trends Microbiol.* 24, 327–337. doi: 10.1016/j.tim.2016.01.008

Wolanin, P. M., Thomason, P. A., and Stock, J. B. (2002). Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol.* 3:REVIEWS3013. doi: 10.1186/gb-2002-3-10-reviews3013

Zankari, E. (2014). Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in bacteria. *Antimicrob. Agents Chemother.* 58:4986. doi: 10.1128/AAC.02620-14

Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinform.* Chapter11:Unit11.5. doi: 10.1002/0471250953.bi1105s31

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comp. Biol.* 7, 203–214. doi: 10.1089/10665270050081478

Zlosnik, J. E., Costa, P. S., Brant, R., Mori, P. Y., Hird, T. J., Fraenkel, M. C., et al. (2011). Mucoid and nonmucoid *Burkholderia cepacia* complex bacteria

in cystic fibrosis infections. *Am. J. Respir. Crit. Care Med.* 183, 67–72. doi: 10.1164/rccm.201002-0203OC

Zlosnik, J. E., Mori, P. Y., To, D., Leung, J., Hird, T. J., and Speert, D. P. (2014). Swimming motility in a longitudinal collection of clinical isolates of *Burkholderia cepacia* complex bacteria from people with cystic fibrosis. *PLoS One* 9:e106428. doi: 10.1371/journal.pone.0106428

Zlosnik, J. E., and Speert, D. P. (2010). The role of mucoidy in virulence of bacteria from the *Burkholderia cepacia* complex: a systematic proteomic and transcriptomic analysis. *J. Infect. Dis.* 202, 770–781. doi: 10.1086/655663

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.

Copyright © 2020 Hassan, dos Santos, Cooper and Sá-Correia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 October 2020
doi: 10.3389/fgene.2020.573395

Characterization of the Complete Mitochondrial Genomes of Two Sibling Species of Parasitic Roundworms, *Haemonchus contortus* and *Teladorsagia circumcincta*

Nikola Palevich[1]*, *Paul H. Maclean*[1], *Young-Jun Choi*[2] and *Makedonka Mitreva*[2]

[1] AgResearch Limited, Grasslands Research Centre, Palmerston North, New Zealand, [2] McDonnell Genome Institute and Department of Medicine, Washington University School of Medicine, Saint Louis, MO, United States

OPEN ACCESS

Edited by:
Jianying Gu,
City University of New York,
United States

Reviewed by:
John Stuart Gilleard,
University of Calgary, Canada
Chun-Ren Wang,
Heilongjiang Bayi Agricultural University, China

***Correspondence:**
Nikola Palevich
nik.palevich@agresearch.co.nz

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

Received: 16 June 2020
Accepted: 18 August 2020
Published: 08 October 2020

Citation:
Palevich N, Maclean PH, Choi Y-J and Mitreva M (2020) Characterization of the Complete Mitochondrial Genomes of Two Sibling Species of Parasitic Roundworms, Haemonchus contortus and Teladorsagia circumcincta. Front. Genet. 11:573395. doi: 10.3389/fgene.2020.573395

Haemonchus contortus and *Teladorsagia circumcincta* are among the two most pathogenic internal parasitic nematodes infecting small ruminants, such as sheep and goats, and are a global animal health issue. Accurate identification and delineation of Haemonchidae species is essential for development of diagnostic and control strategies with high resolution for Trichostrongyloidea infection in ruminants. Here, we describe in detail and compare the complete mitochondrial (mt) genomes of the New Zealand *H. contortus* and *T. circumcincta* field strains to improve our understanding of species- and strain-level evolution in these closely related roundworms. In the present study, we performed extensive comparative bioinformatics analyses on the recently sequenced complete mt genomes of the New Zealand *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP field strains. Amino acid sequences inferred from individual genes of each of the two mt genomes were compared, concatenated and subjected to phylogenetic analysis using Bayesian inference (BI), Maximum Likelihood (ML), and Maximum Parsimony (MP). The AT-rich mt genomes of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP are 14,001 bp (A+T content of 77.4%) and 14,081 bp (A+T content of 77.3%) in size, respectively. All 36 of the typical nematode mt genes are transcribed in the forward direction in both species and comprise of 12 protein-encoding genes (PCGs), 2 ribosomal RNA (*rrn*) genes, and 22 transfer RNA (*trn*) genes. The secondary structures for the 22 *trn* genes and two *rrn* genes differ between *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP, however the gene arrangements of both are consistent with other Trichostrongylidea sequenced to date. Comparative analyses of the complete mitochondrial nucleotide sequences, PCGs, A+T rich and non-coding repeat regions of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP further reinforces the high levels of diversity and gene flow observed among Trichostrongylidea, and supports their potential as ideal markers for strain-level identification from different

hosts and geographical regions with high resolution for future studies. The complete mt genomes of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP presented here provide useful novel markers for further studies of the meta-population connectivity and the genetic mechanisms driving evolution in nematode species.

Keywords: *Haemonchus contortus*, *Teladorsagia circumcincta*, mitochondrial genome, helminth, parasite, anthelmintic-susceptible

HIGHLIGHTS

- In-depth comparative and phylogenomic analyses of the recently published complete mitochondrial (mt) genomes of the New Zealand *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP field strains to improve our understanding of species- and strain-level evolution in roundworms.
- Analysis of the A+T rich and non-coding repeat regions of the complete *Haemonchus* and *T. circumcincta* mt genomes supports their potential as ideal markers for easy species- and possibly strain-level identification with high resolution for future studies.
- Phylogenetic relationships of complete mt genomes for all Trichostrongyloidea species currently available inferred from Bayesian, Maximum Likelihood and Maximum Parsimony analyses of mitochondrial genes, revealed high levels of intra-species diversity and a panmictic structure observed among Trichostrongylidea.
- Our findings indicate that a three-pronged approach incorporating phylogenetic inertia, pangenome structure/features and environmental data are needed in order to understand the mitochondrial genome evolution.

INTRODUCTION

H. contortus (barber's pole worm) and *T. circumcincta* (brown stomach worm), are the most economically important pathogenic nematodes infecting small ruminants (sheep and goats) worldwide. These blood-feeding strongylid nematodes are orally transmitted via contaminated pasture to the host where they infect the fourth stomach (abomasum) reducing animal production and causing anemia, edema, and associated complications often leading to death (Vlassoff et al., 2001; Sutherland and Scott, 2010). Although these parasites can be managed using existing prophylactic drugs (anthelmintics), their remarkable natural tendency to develop resistance combined with the diminishing efficacy of compounds used, threatens the global livestock industry (Kaplan, 2004; Kaplan and Vidyashankar, 2012). While these monoxenous and obligately sexual species have identical life-cycles (Veglia, 1915), the gene flow and evolutionary relationships observed among the Trichostrongylidae superfamily, and the Strongylida suborder as a whole, remain unclear.

Parasitic roundworms belonging to the Haemonchidae family are able to infect a range of ruminant hosts. For example and on communal pastures in particular, *H. placei* is primarily a cattle parasite but can also infect small ruminants (Amarante et al., 1997; Gasser et al., 2008; Nunes et al., 2013), and vice versa for *H. contortus* (Jacquiet et al., 1998). Given the increase in availability of genomic information resources (Palevich et al., 2018), it is crucial for the management of parasitic diseases that readily accessible molecular tools are developed to enable end-users to easily determine which species is infecting which animal host.

The development of low cost, high-throughput, next-generation sequencing technologies have led to a recent focus on whole-genome sequencing (WGS) of complete nuclear genomes of parasitic roundworms (Nematodes) (Hu and Gasser, 2006; Hu et al., 2007; Jex et al., 2008; Palevich et al., 2019a,b), instead of earlier markers such as *cox1*, *cox2*, and *nad* genes. Owing to the substantial coverage following WGS (Wit and Gilleard, 2017; Palevich et al., 2019c) and advances in bioinformatics pipelines, further complete mitochondrial genome resources will be available to facilitate detailed comparative phylogenetic analyses across many species complexes and major taxonomic groups of nematodes (Hu et al., 2004; Gasser et al., 2016). In general, complete nematode mt genomes are ∼12–21 kbp circular-DNA molecules (Hu and Gasser, 2006). Nematode mt genomes can provide rich sources of informative markers and typically feature (Wolstenholme, 1992; Boore, 1999; Saccone et al., 1999; Blouin, 2002): 12 protein-coding genes (PCGs) and lack of the *atp8* gene [except for *T. spiralis* (Lavrov and Brown, 2001)], 2 ribosomal RNA (*rrn*) genes (Gutell et al., 1993), and 22 transfer RNA (*trn*) genes (Wolstenholme et al., 1987); unidirectional transcription (usually in forward direction) with particularly unique initiation codons; and highly variable gene arrangements (Hu and Gasser, 2006). The gene sequences, and in particular the 12 PCGs, are relatively conserved are useful phylogenetic markers that have been used for examining phylogenetic relationships within the phylum Nematoda (Hu et al., 2004). The mitochondrial genetic code of nematodes follows translation (transl_table = 5)[1] of National Center for Biotechnology Information (NCBI). In addition, nematode mt genomes also possess transposition genotypes of tRNAs that lack either a TΨC or DHU stem in the *trn* secondary structures (Wolstenholme et al., 1994; Yokobori, 1995). In the pursuit of improving the phylogenetic resolution within the phylum Nematoda, mt genomes serve as excellent markers for investigation of species delineation and the genetics of population evolution. Future efforts should focus on the availability of more complete mt genomes across all nematode species, and especially for conspecific strains.

For these reasons, this study interrogates the complete mitochondrial genome sequences of the anthelmintic-susceptible

[1] https://ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?chapter=tgencodes#SG5

H. contortus NZ_Hco_NP (Palevich et al., 2019a) and *T. circumcincta* NZ_Teci_NP (Palevich et al., 2019b) derived from pasture-grazed New Zealand sheep. In this study, we performed numerous comparative and phylogenomic analyses to directly compare the molecular characteristics including the nucleotide composition and codon usage profiles of PCGs, as well as the secondary structures of each identified tRNA and rRNA gene within the two species, but also among other closely related trichostrongyloid nematodes. The phylogenetic position of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP among Trichostrongylidae and of other species of socio-economically important parasites within the phylum Nematoda, are investigated based on mitochondrial nucleotide sequences and PCGs. This research provides insights into the mt genome evolution that may facilitate the development of molecular tools to differentiate between these two sibling species of parasitic roundworms at the strain level.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

The acquisition of parasite samples, DNA extraction, and genome sequencing technologies used to generate the mt genomes used for analysis in this study have recently been outlined in the form of brief reports (Palevich et al., 2019a,b). To fill any knowledge gaps associated with the above-mentioned reports, here we describe in detail the methodologies, sequencing technologies, and annotation software used to generate the *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mt genomes. *H. contortus* and *T. circumcincta* were recovered from sheep infected with pure strains of parasites in Palmerston North, New Zealand, and total high molecular weight genomic DNA was extracted separately from multiple worms using a modified phenol:chloroform protocol, as previously described (Choi et al., 2017; Palevich et al., 2019a,b,c). DNA was deposited, stored and available upon request from AgResearch Ltd., Grasslands Research Center.

Mitochondrial Genome Sequencing and Annotation

The complete mitochondrial genome sequences of *H. contortus* NZ_Hco_NP (Palevich et al., 2019a,c) and *T. circumcincta* NZ_Teci_NP (Palevich et al., 2019b) were obtained using next-generation sequencing technologies. The *H. contortus* NZ_Hco_NP (BioProject ID: PRJNA517503, SRA accession number SRP247265) whole genome shotgun paired-end (PE) and Single-Molecule, Real-Time (SMRT) long-read sequencing (SRA Runs: SRR11022845 and SRR11022846) were generated using the Illumina HiSeq2500 and Pacific Biosciences (PacBio) platforms, as previously described (Palevich et al., 2019a,c). The *T. circumcincta* NZ_Teci_NP (BioProject ID: PRJNA72569, SRA accession number SRP007648) whole genome shotgun library (SRA Run: SRR3145390) was generated as previously described (Tang et al., 2014; Palevich et al., 2019b), and sequenced by the Genome Center at Washington University (WUGSC, School of Medicine, St. Louis, MO, USA) using the Illumina MiSeq platform with the strategy of 150 bp paired-end sequencing mode. A total of 605 million raw reads were generated and made available in FASTQ format. The quality of the raw sequence reads was evaluated using the software package [http://www.bioinformatics.babraham.ac.uk/projects/fastqc (Andrews, 2010)]. The software Trimmomatic v.0.36 (Bolger et al., 2014) was used for removal of adapter, contaminant, low quality (Phred scores <30), and short (<50 bp) sequencing reads. The remaining high-quality sequencing reads were assembled *de novo* using the NOVOPlasty pipeline v.3.1 with default parameters and based on a kmer size of 39 following the developer's suggestions (Dierckxsens et al., 2016). The assembled mt genomes, tRNA and rRNA annotations were checked using the MITOS web server based on translation (transl_table = 5)[1] (Invertebrate Mitochondrial) of NCBI (http://mitos.bioinf.uni-leipzig.de). Repeat-rich subsets of the mt genome assemblies were identified with varying degrees of complexity using the mDUST v.2006.10.17 (Morgulis et al., 2006) software tools with default settings. The sequence regions containing tandem repetitive elements were identified using Tandem Repeat Finder software online server (http://tandem.bu.edu/trf/trf.html). The coordinates of the found repeats were matched with coding sequences and the proportion of repeats overlapping coding regions was calculated. tRNA genes were further corroborated using the software SPOT-RNA (Singh et al., 2019) and the tRNA secondary structures were predicted using the ViennaRNA v.2.4 (Lorenz et al., 2011) and VARNA v.3.93 (Darty et al., 2009) packages. tRNA secondary structures and complete mt genomes were visualized using the Geneious Prime v.2019.1.1 (Kearse et al., 2012). The PCGs were manually curated by searching for ORFs (employing genetic code 5 as above) and alignment against the available *H. contortus* and *T. circumcincta* reference genomes in Geneious Prime v.2019.1.1 (Kearse et al., 2012).

Mitochondrial Pangenome Analysis

The entire nucleotide sequences alignments using Mauve (Darling et al., 2004) were performed using Geneious Prime v.2019.1.1 (Kearse et al., 2012). Sequence differences among the *Haemonchus* and *Teladorsagia* species as well as between the respective strains was done using the *H. contortus* NZ_Hco_NP (Palevich et al., 2019a,c) and *T. circumcincta* NZ_Teci_NP (Choi et al., 2017; Palevich et al., 2019b) as the reference sequences, respectively. Repeated sequence elements of *Haemonchus* and *Teladorsagia* mt genomes were visualized in the form of Dotplots and generated based on pair-wise alignments with chaining coverage using PipMaker (Schwartz et al., 2000). Gene synteny was investigated using BLASTX alignments of sequences and genetic similarity mapping was visualized using the CGView Comparison Tool (http://stothard.afns.ualberta.ca/downloads/CCT) (Grant et al., 2012). For analysis of codon usage and amino acid composition, codon usage counts were extracted from each protein using Geneious Prime version 2019.1.1 (Kearse et al., 2012). These were then compiled and loaded into R version v.3.6.1 (R Core Team, 2013) where a clustered image map was produced using the "cim" function from the mixOmics package v.6.8.5 (Rohart et al., 2017).

Phylogenomics Analysis of Complete mt Genomes

The phylogenetic positions of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP among other species of nematodes available in the GenBank database were examined. The nucleotide sequences and PCGs of published Trichostrongyloidea mitochondrial genomes were retrieved from the National Center for Biotechnology Information, USA (http://www.ncbi.nlm.nih.gov/). *Caenorhabditis elegans* N2 (JF896456) was used as the outgroup species in this study. All coding sequences (CDS) used in this analysis had correct start and stop codons based on translation (transl_table = 5)[1] of NCBI (Invertebrate Mitochondrial).

Considering the high degree of intraspecific variation in nucleotide sequences of mt genes of nematodes, we used the deduced amino acid sequences of mt proteins for phylogenetic analyses. Amino acid sequences from the 12 mt protein-coding genes of the complete mitochondrial genomes were deduced, aligned individually and concatenated into a single alignment using MAFFT v.7.450 (Katoh and Standley, 2013). Phylogenetic analyses were conducted using three methods: Bayesian inference (BI), Maximum Likelihood (ML), and Maximum Parsimony (MP) methods. BI analysis with the program MrBayes v.3.15265 (Huelsenbeck and Ronquist, 2001) and the MtInv model of amino acid evolution (Le et al., 2017) was selected as the most suitable model of evolution. However, since the MtInv model is not implemented in the current version of MrBayes, an alternative model, MtREV (Adachi and Hasegawa, 1996), was used in BI analysis. Four independent Markov chains were run for 1,000,000 metropolis-coupled MCMC generations, sampling a tree every 100 generations. The first 2,500 trees represented burn-in, and the remaining trees were used to calculate Bayesian posterior probabilities (PP). The evolutionary history was further inferred by using the ML and MP methods. ML analysis was performed with the MtREV General Reversible Mitochondrial + Freq. model (Adachi and Hasegawa, 1996) and MP was performed with PAUP* v.4.0b10 (http://paup.csit.fsu.edu/) (Swofford, 2001). Both methods were calculated using 1,000 bootstrap replicates with bootstrapping frequencies (BF) were calculated. Phylograms were drawn using the software MEGA X (Kumar et al., 2018).

RESULTS AND DISCUSSION

mt Genome General Features and Characteristics

The reconstruction of partial and complete mitochondrial genomes is becoming increasingly cost-effective with the advent of next-generation sequencing technology and recent advances in molecular parasitology tools, in particular for species identification and population genetics studies (Hu and Gasser, 2006; Jex et al., 2009). Despite these advances, only 12 species of Trichostrongyloid mt genomes are currently available in GenBank with only half of these represented by multiple strains (dos Santos et al., 2017). For these reasons, the complete mt genomes of New Zealand *H. contortus* NZ_Hco_NP (Palevich et al., 2019a,c) and *T. circumcincta* NZ_Teci_NP (Choi et al., 2017; Palevich et al., 2019b) field strains were sequenced and compared in detail to investigate the strain-level genetic differences of these two sibling Trichostrongyloidea species. Recently, complete mt genomes of NZ_Hco_NP and NZ_Teci_NP were generated using various sequencing platforms and assembled from 582 and 5.6 million (PE) reads with an average coverage depth per nucleotide of 13,293× and 2,097×, respectively.

The smallest Trichostrongyloidea mt genome published to date belongs to *Dictyocaulus* sp. cf. *eckerti* (Gasser et al., 2012) at 13,296 bp with the largest 15,221 bp mt genome belonging to *M. digitatus* (Jex et al., 2009). The NZ_Hco_NP and NZ_Teci_NP mt genomes are 14,001 and 14,081 bp in length, respectively (**Figure 1**) and are thus within the expected range. The reported mt genomes differ in size by ∼16–54 bp than other *H. contortus* and *T. circumcincta* characterized strains, except for the 13.7 kb *H. placei* MHpl1 mt genome (dos Santos et al., 2017). The size discrepancies among the different strains primarily relate to A+T-rich control region expansions and longer non-coding regions between numerous transfer RNA genes.

The total GC content of the *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mt genomes were 21.1 and 22.8% and contained an A+T bias with an overall base composition of A = 32.9 and 31.2%, T = 44.5 and 46.1%, C = 6.3 and 7.3%, and G = 14.8 and 15.5%, respectively (**Table 1**). The overall A+T bias of 77.4 and 77.3% observed in *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mt genomes, respectively, are within the typical range reported for nematode mitochondrial genomes (Hu et al., 2004; Hu and Gasser, 2006). Compared with the others, the A+T content of all *T. circumcincta* strains are within 0.1% of each other with *H. contortus* strains being ±0.5 of NZ_Hco_NP. Low-complexity regions (LCRs) are extremely abundant in eukaryotic proteins that are typically represented by amino acid sequences containing repeats of single amino acids or short amino acid motifs, and may facilitate the formation of novel coding sequences. *H. contortus* NZ_Hco_NP had 23 LCRs with 15 (65%) in coding regions, while *T. circumcincta* NZ_Teci_NP had 26 repeats with 21 (81%) found in coding regions (**Figure 2**), however determining the functional role of these repeats is a focus for future work. Comparative analysis of the *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mtDNAs also revealed five and four tandem repeat (TR) regions, respectively (**Figure 2**).

To date, mt genomes are available for many species of Trichostrongyloidea and the mt genomes of the trichostrongyloid *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP are in agreement, as they both comprise of 12 protein-encoding genes (PCGs), 22 transfer RNA (*trn*) genes, and 2 ribosomal RNA genes (*rrnS* or 12*S* ribosomal RNA and *rrnL* or 16*S* ribosomal RNA). The 12 mt PCGs of both *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP were encoded on the heavy strand and transcribed in the same direction. The order of genes observed in *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP are identical to the previously reported *H. contortus* (Palevich et al., 2018) and *T. circumcincta* (Jex et al., 2009), respectively (**Figure 1**). In accordance to the mt genomes of

TABLE 1 | Comparison of the annotated mitochondrial genomes of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP.

Name	Type	Position		Length (bp)	Strand	Codons of PCGs		Nucleotide composition							Nucleotide skewness	
		Start	Stop			Start	Stop	T (%)	C (%)	A (%)	G (%)	AT (%)	GC (%)	GT (%)	AT skew	GC skew
Protein-encoding genes (PCGs)																
*atp*6	Coding	8,809|8,901	9,408|9,500	600|600	+	ATT|ATT	TAA|TAA	46.3|46.7	6.3|6.7	29.5|30.8	16.3|15.8	75.8|77.5	22.6|22.5	62.6|62.5	−0.222|−0.205	0.442|0.404
*cox*1	Coding	1|7	1,582|1,578	1,582|1,572	+	ATA|ATA	TAA|TAA	42.7|43.8	9.7|10.9	26.5|25.1	19.6|20.3	69.2|68.9	29.3|31.2	62.3|64.1	−0.234|−0.271	0.338|0.301
*cox*2	Coding	1,862|1,898	2,554|2,593	693|696	+	ATA|ATT	TAG|TAG	42.6|43.7	7.8|9.8	30.4|28.0	17.5|18.5	73.0|71.7	25.3|28.3	60.1|62.2	−0.167|−0.219	0.383|0.307
*cox*3	Coding	11,880|11,959	12,648|12,759	769|801	+	ATA|ATT	TAA|TAA	46.8|46.1	7.4|7.3	24.7|31.2	17.4|15.5	71.5|77.3	24.8|22.8	64.2|61.6	−0.309|−0.193	0.403|0.36
*cyt*b	Coding	10,711|10,790	11,823|11,902	1,113|1,113	+	ATT|ATA	TAA|TAA	45.7|45.6	7.4|8.6	29.5|28.5	15.5|17.3	75.2|74.1	22.9|25.9	61.2|62.9	−0.215|−0.231	0.354|0.336
*nad*1	Coding	7,935|8,030	8,807|8,905	873|876	+	ATA|ATT	TAA|TAA	45.6|48.6	7.9|7.6	26.2|25.0	19.8|18.7	71.8|73.6	27.7|26.3	65.4|67.3	−0.27|−0.321	0.43|0.422
*nad*2	Coding	9,598|9,705	10,443|10,544	846|840	+	TTG|ATT	TAA|TAA	47.8|51.1	4.4|5.5	35.7|31.2	11.9|12.3	83.5|82.3	16.3|17.8	59.7|63.4	−0.145|−0.242	0.46|0.382
*nad*3	Coding	3,559|3,598	3,894|3,933	336|336	+	TTG|ATT	TAA|TAA	49.7|47.0	2.4|3.6	30.4|31.8	15.8|17.6	80.1|78.8	18.2|21.2	65.5|64.6	−0.241|−0.193	0.736|0.66
*nad*4	Coding	12,701|12,796	13,930|14,007	1,230|1,212	+	ATA|ATT	TAA|TAA	48.1|49.1	7.6|8.4	30.2|28.6	11.1|13.9	78.3|77.7	18.7|22.3	59.2|63.0	−0.229|−0.264	0.187|0.247
*nad*4L	Coding	6,688|6,672	9,919|6,903	232|232	+	ATT|ATT	TAA|TAA	53.0|57.3	2.6|2.2	29.3|26.7	13.8|13.8	82.3|84.0	16.4|16.0	66.8|71.1	−0.288|−0.364	0.683|0.725
*nad*5	Coding	3,896|3,942	5,477|5,520	1,582|1,579	+	ATT|ATT	TAA|TAA	44.5|48.8	6.3|7.3	32.9|30.0	14.8|13.9	77.4|78.8	21.1|21.2	59.3|62.7	−0.15|−0.239	0.403|0.311
*nad*6	Coding	6,241|6,135	6,681|6,572	441|438	+	ATT|ATT	TAA|TAA	46.3|52.5	4.3|6.2	29.7|27.4	19.3|13.9	76.0|79.9	23.6|20.1	65.6|66.4	−0.218|−0.314	0.636|0.383
Transfer RNA																
trnA	tRNA-Ala	5,478|5,521	5,534|5,578	57|58	+	GGG|GGG	CTA|TAA	42.1|37.9	7|8.6	40.4|37.9	10.5|15.5	82.5|75.8	17.5|24.1	52.6|53.4	−0.021|0	0.2|0.286
trnR	tRNA-Arg	10,514|10,620	10,569|10,674	56|55	+	AAA|AAA	TTT|TTT	37.5|43.6	8.9|7.3	41.1|36.4	12.5|12.7	78.6|80.0	21.4|20.0	50.0|56.3	0.046|−0.09	0.168|0.27
trnN	tRNA-Asn	7,821|7,879	7,876|7,938	56|60	+	TAA|TTA	TAA|AGT	41.1|41.7	5.4|6.7	35.7|31.7	14.3|20	76.8|73.4	19.7|26.7	55.4|61.7	−0.07|−0.136	0.452|0.498
trnD	tRNA-Asp	1,728|1,782	1,784|1,839	57|58	+	AAA|AAA	TAA|TAG	36.8|39.7	3.5|3.4	45.6|41.4	14|15.5	82.4|81.1	17.5|18.9	50.8|55.2	0.107|0.021	0.6|0.64
trnC	tRNA-Cys	1,583|1,579	1,636|1,633	54|55	+	ATT|ATT	ATT|TTT	40.7|45.5	3.7|5.5	46.3|40	9.3|9.1	87.0|85.5	13.0|14.6	50.0|54.6	0.064|−0.064	0.431|0.247
trnQ	tRNA-Gln	10,570|10,678	10,624|10,733	55|56	+	TGT|TTG	CAG|AAA	43.6|42.9	1.8|1.8	36.4|39.3	16.4|16.1	80.0|82.2	18.2|17.9	60.0|59.0	−0.09|−0.044	0.802|0.799
trnE	tRNA-Glu	6,988|6,990	7,042|7,047	55|58	+	GAG|GAG	TTG|TGT	38.2|41.4	3.6|3.4	43.6|37.9	14.5|17.2	81.8|79.3	18.1|20.6	52.7|58.6	0.066|−0.044	0.602|0.67
trnG	tRNA-Gly	1,807|1,845	1,861|1,898	55|54	+	AAT|ATT	TTA|TAA	41.8|42.6	3.6|5.6	45.5|37.0	7.3|14.8	87.3|79.6	10.9|20.4	49.1|57.4	0.042|−0.07	0.339|0.451
trnH	tRNA-His	2,556|2,592	2,610|2,646	55|55	+	AGC|AGC	CTA|TAT	32.7|38.2	3.6|5.5	43.6|36.4	16.4|20	76.3|74.6	20.0|25.5	49.1|58.2	0.143|−0.024	0.64|0.569
trnI	tRNA-Ile	10,455|10,552	10,513|10,612	59|61	+	ATT|ATT	ATA|TAG	39|39.3	6.8|4.9	42.4|42.6	11.9|13.1	81.4|81.9	18.7|18.0	50.9|52.4	0.042|0.04	0.273|0.456
trnL1	tRNA-Leu	9,489|9,596	9,543|9,652	55|57	+	GTT|GTT	ACT|CTA	29.1|33.3	9.1|7	43.6|43.9	18.2|15.8	72.7|77.2	27.3|22.8	47.3|49.1	0.199|0.137	0.333|0.386
trnL2	tRNA-Leu	11,824|11,904	11,879|11,958	56|55	+	GCA|TAC	GCT|ATT	33.9|41.8	5.4|3.6	46.4|43.6	8.9|10.9	80.3|85.4	14.3|14.5	42.8|52.7	0.156|0.021	0.245|0.503
trnK	tRNA-Lys	9,423|9,499	9,480|9,561	58|63	+	GTT|AAA	TTA|TAT	41.4|39.7	8.6|9.5	37.9|39.7	10.3|11.1	79.3|79.4	18.9|20.6	51.7|50.8	−0.044|0	0.09|0.078
trnM	tRNA-Met	1,649|1,683	1,705|1,742	57|60	+	AAT|AGT	TTA|TAT	33.3|31.7	10.5|13.3	35.1|33.3	21.1|21.7	68.4|65.0	31.6|35.0	54.4|53.4	0.026|0.025	0.335|0.24
trnF	tRNA-Phe	10,657|10,736	10,710|10,790	54|55	+	ATT|ATC	ATA|TAA	40.7|32.7	0.0|5.5	50|50.9	9.3|10.9	90.7|83.6	9.3|16.4	50.0|43.6	0.103|0.218	1|0.329
trnP	tRNA-Pro	6,085|5,932	6,140|5,988	56|57	+	CAA|CAA	TGA|GAA	35.7|42.1	5.4|1.8	39.3|43.9	19.6|12.3	75.0|86.0	25.0|14.1	55.3|54.4	0.048|0.021	0.568|0.745
trnS1	tRNA-Ser	7,744|7,748	7,798|7,803	55|56	+	AAA|TAA	TTT|TAT	41.8|39.3	3.6|7.1	43.6|39.3	9.1|14.3	85.4|78.6	12.7|21.4	50.9|53.6	0.021|0	0.433|0.336
trnS2	tRNA-Ser	9,544|9,652	9,597|9,705	54|54	+	AAC|AAT	TTT|TTA	38.9|44.4	13|13	29.6|25.9	16.7|16.7	68.5|70.3	29.7|29.7	55.6|61.1	−0.136|−0.263	0.125|0.125
trnT	tRNA-Thr	12,647|12,724	12,702|12,777	56|54	+	GTT|TGT	ACT|AAC	44.6|42.6	3.6|5.6	44.6|37	7.1|14.8	89.2|79.6	10.7|20.4	51.7|57.4	0|−0.07	0.327|0.451
trnW	tRNA-Trp	6,920|6,904	6,976|6,961	57|58	+	GTA|ATA	ATA|TAT	42.1|41.4	5.3|5.2	45.6|48.3	7|5.2	87.7|89.7	12.3|10.4	49.1|46.6	0.04|0.077	0.138|0
trnY	tRNA-Tyr	7,881|7,975	7,934|8,030	54|56	+	AAG|AAG	TTA|TGA	42.6|42.9	1.9|1.8	42.6|41.1	11.1|14.3	85.2|84.0	13.0|16.1	53.7|57.2	0|−0.021	0.708|0.776
trnV	tRNA-Val	6,180|6,081	6,234|6,135	55|55	+	AAA|AAT	TTT|CTA	41.8|40	1.8|5.5	49.1|47.3	7.3|7.3	90.9|87.3	9.1|12.8	49.1|47.3	0.08|0.084	0.604|0.141
Ribosomal RNA																
rrnS	Small subunit rRNA (12*S*)	7,041|7,050	7,743|7,749	703|700	+	TGT|AAG	TAA|TTA	38.0|41.0	5.8|7.4	41.8|36.0	13.5|15.6	79.8|77.0	19.3|23.0	51.5|56.6	0.048|−0.065	0.399|0.357
rrnL	Large subunit rRNA (16*S*)	2,611|2,638	3,561|3,599	951|962	+	TAA|AAA	TTG|CAT	40.7|43.3	5.3|5.5	42.2|39.7	10.7|11.4	82.9|83.0	16.0|16.9	51.4|54.7	0.018|−0.043	0.338|0.349
Other features																
AT-rich region	Non-coding	5,535	6,118	583	+	GAA	TTG	44.6	2.4	45.5	6.5	90.1	8.9	51.1	0.010	0.461
Full genome				14,001|14,081				44.5|46.1	6.3|7.3	32.9|31.2	14.8|15.5	77.4|77.3	21.1|22.8	59.3|61.6	−0.15|−0.193	0.403|0.36

H. contortus NZ_Hco_NP|T. circumcincta NZ_Teci_NP.

other Trichostrongyloidea, the reported mt genomes also all lack the *atp*8 gene. All of the 12 PCGs in both mt genomes are predicted to use unique translation initiation codons such as ATT (*atp*6, *cyt*b, and *nad*4-6), TTG (*nad*2-3), and ATA (*nad*1 and *cox*1-3) in *H. contortus* NZ_Hco_NP, with *T. circumcincta* NZ_Teci_NP predominantly using ATT and ATA (*cyt*b and *cox*1) (**Table 1**). Both mt genomes are predicted to use the complete termination codons TAA. Nevertheless, *cox*2 had a TAG translation termination codon similar to *nad*3 of *H. placei* where not all termination codons were complete stop codons, typical of nematodes (Jex et al., 2008; dos Santos et al., 2017).

Of the 22 predicted *trn* genes, size ranged from 54 to 63 bp in length (**Table 1**), in *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mt genome sequences and are similar to those of other nematodes studied to date (**Supplementary Figure 1**). The characteristics associated with the secondary structures predicted for most tRNA genes include: a 7–8 bp acceptor stem (amino-acyl arm), a 5 bp anticodon stem and a T/U residue preceding. With the exception of *T. spiralis* (Lavrov and Brown, 2001), these characteristics are consistent with all previously published mt genomes for the Trichostrongyloidea (dos Santos et al., 2017) and Rhabditida [*C. elegans* (Blaxter et al., 1998)].

The *rrnS* (12*S*) and *rrnL* (16*S*) genes identified in the mt genomes of *H. contortus* NZ_Hco_NP (703 and 951 bp) and *T. circumcincta* NZ_Teci_NP (700 and 962 bp) were A+T biased (**Table 1**). Secondary structure models for the *rrn* genes (**Supplementary Figure 2**) of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP were constructed based on models of nematodes that have been previously reported (Hu et al., 2004, 2007; Hu and Gasser, 2006). The *rrn* genes were particularly A+T rich with 79.8 and 77.0% for *rrnS* and 82.9 and 83.0% for *rrnL* in *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP, respectively. The overall base compositions of the *rrnS* genes were; A = 41.8 and 36.0%, T = 38.0 and 41.0%, C = 5.8 and 7.4%, and G = 13.5 and 15.6%, and that of the *rrnL* genes were; A = 42.2 and 39.7%, T = 40.7 and 43.3%, C = 5.3 and 5.5%, and G = 10.7 and 11.4%. Therefore, the A+T content of the *rrnS* and *rrnL* genes were comparable to the overall A+T content of both whole mt genomes. The *rrnL* gene in both *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP, is located between *trnH* and *nad*3. Similarly, the

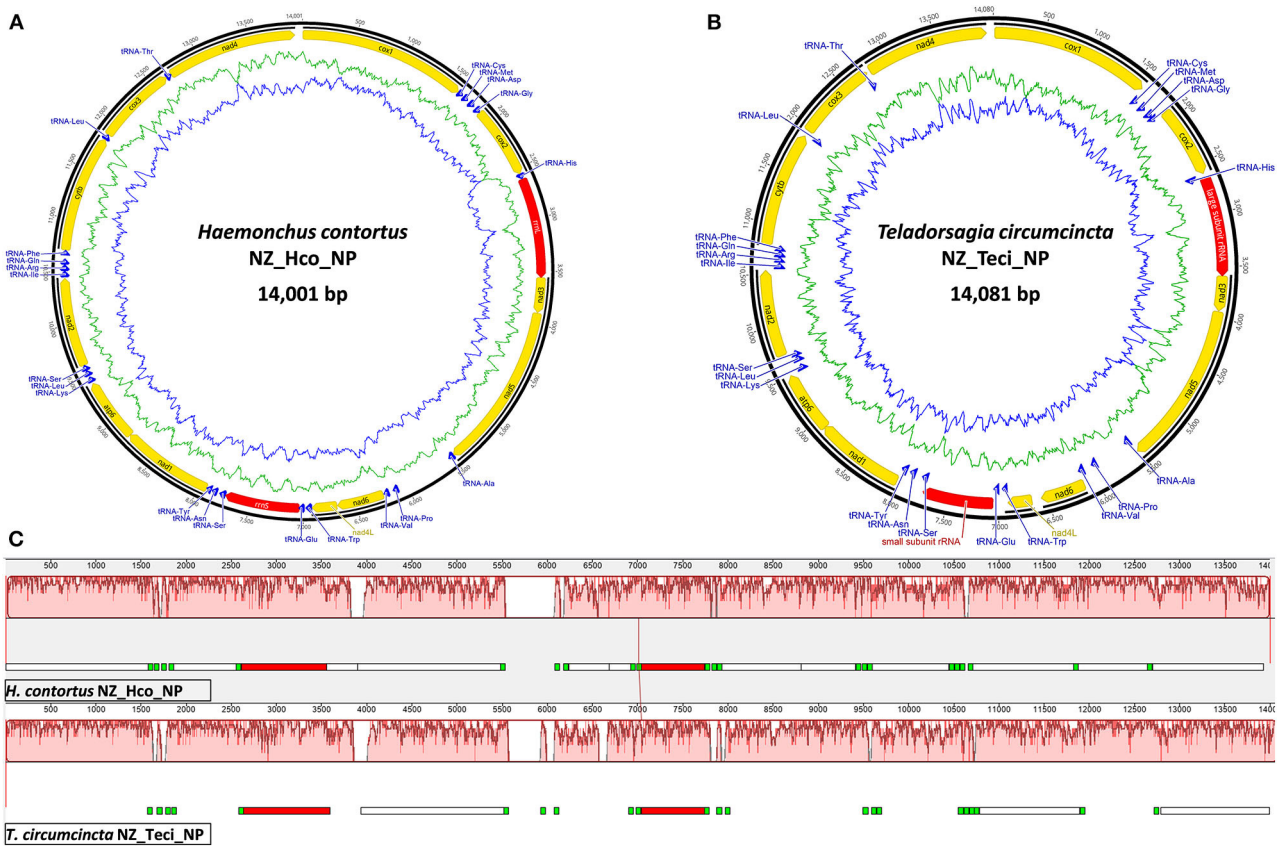


FIGURE 1 | Mitochondrial genome atlases of *H. contortus* NZ_Hco_NP **(A)** and *T. circumcincta* NZ_Teci_NP **(B)**. Each map is annotated and depicts the 12 protein-coding genes (PCGs, yellow), two ribosomal RNA genes (*rrnS* and *rrnL*, red), 22 transfer RNA (*trn*, blue) genes, and any putative non-coding region if applicable (gray). The innermost circles depict GC (blue) and AT (green) content, respectively, along the genome. Mauve visualization of an alignment of the *H. contortus* NZ_Hco_NP (top) and *T. circumcincta* NZ_Teci_NP (bottom) complete mt genomes **(C)**. Mauve alignments of each genome are represented by a horizontal track, with annotated coding regions (white boxes), *trn* genes (green), and *rrn* genes (red). Red colored segments represent conserved regions among the two mt genomes.

rrnS gene is located between *trnE* and *trnS1* almost 3.5 Kb from the *rrnL* and between *nad*4L and *nad*1 in both mt genomes (**Figures 1**, **2**). The locations, sizes and secondary structure characteristics of the *rrnL* and *rrnS* genes are relatively conserved among all Trichostrongyloidea mt genomes analyzed, especially the stem regions of the *rrnL* genes between *H. contortus* and *T. circumcincta* strains.

Comparative mt Pangenome Analysis of Trichostrongyloidea

Comparisons of the gene synteny as well as the nucleotide, codon usage and amino acid compositions revealed a highly conserved GA2 gene arrangement (*nad*6, *nad*4L, *rrnS*, *nad*1, *atp*6, *nad*2, *cyt*b, *cox*3, *nad*4, *cox*1, *cox*2, *rrnL*, *nad*3, *nad*5), and mt genome structure as previously reported for trichostrongyloids (Hu et al., 2004, 2007). Also, in accordance with previously published mt genomes the absence of the presumed non-essential ATP synthase F0 subunit 8 (*atp*8) gene reported only in *Trichinella spiralis* (Lavrov and Brown, 2001; Jex et al., 2008, 2009) and other *Trichuris* species such as *Trichuris ovis* and *Trichuris discolor* (Nematoda: Trichuridae), was absent in our data.

The predicted lengths of all the genes in *H. contortus* NZ_Hco_NP are very similar to those of *T. circumcincta* NZ_Teci_NP ($\leq$1 amino acid difference). The sequence variability between *Haemonchus* species/strains and *T. circumcincta* strains of the Trichostrongyloidea superfamily was assessed by performing pairwise comparisons for each PCG of the mt genomes (**Figure 2**). Among the *Haemonchus* species/strains, the most variable genes were *nad*6 and *nad*2 with 13.8 and 10.2% of sequence divergence, respectively, and the most conserved genes were *cox*1, *atp*6, and *cox*2 with 93.6, 92.2, and 92.5% of sequence identity, respectively. For the *T. circumcincta* strains, we found that the most variable gene was *nad4* with 2.2% sequence divergence, with most genes being conserved at around 98.6% sequence identity. Based on our results, the most variable genes among Trichostrongyloidea nematodes were *nad*2, *nad*6, *nad*3, and *nad*5 with 56.1, 38.4, 23.1, and 21.3% of sequence divergence, respectively, with the most conserved genes were *cox*1, *atp*6, *cox*2, and *cox*3 with 87.9, 83.5, 85.3, and 86.9% of sequence identity, respectively. Pairwise comparisons of all the PCGs among Trichostrongyloidea nematodes and it was observed that the *nad4* gene was the most variable in terms of both the gene length and the sequence identity (**Figure 2**), where the predicted lengths differ by exactly 6 amino acids between *Haemonchus* (1,230 bp for *H. contortus* and *H. placei*) and *T. circumcincta* strains (1,212 bp). In comparison, the *nad*4 gene appears to be conserved among different Trichostrongyloidea species, for example *Trichostrongylus vitrinus* and *T. axei* (1,227 bp), while *M. digitatus*, *C. oncophora*, and *C. elegans* N2 all share 1,230 bp *nad*4 genes (Blouin et al., 1997; Blouin, 2002; Hu et al., 2004; Jex et al., 2009; Gasser et al., 2012; Xu et al., 2015).

A recent investigation comparing the divergences observed between ITS-1 and ITS-2 genetic nuclear markers with *cox*1 or *nad*4 genes (Blouin et al., 1997; Blouin, 2002), showed that mt DNA accumulates substitutions more quickly than nuclear genome markers and that *nad*4 genes in particular are a superior tool for prospecting species. For example, regarding *H. placei* and *H. contortus* mt genomes, a divergence in nuclear ITS-2 rDNA between these two species has previously been reported as 1.3% (Stevenson et al., 1995). Whereas, differences between nucleotide sequences of *H. placei* and *H. contortus* mt genome PCGs ranged between 10 and 20% (Gasser et al., 2006; dos Santos et al., 2017). Taken together, our findings are in agreement with previous reports suggesting the use of the most conserved mt *cox* genes for investigating systematics and speciation in nematodes (Hu et al., 2004; Hu and Gasser, 2006; Zarowiecki et al., 2007), while the variation observed in *nad* mt genes would be more appropriate for population genetics studies (Jex et al., 2009).

Analysis of codon usage patterns in Trichostrongyloidea and *C. elegans* N2 revealed that certain codons are predicted to be utilized almost ubiquitously across all mitochondrial PCGs (**Figure 3**). As expected, the A+T bias reported in the Trichostrongyloidea mt genome sequences also affects the amino acid sequence composition of the predicted proteins. Within each codon family, we found that the non-polar Leu (TTG, 5.0%), basic His (CAC, 0.4%), polar Asn (AAC, 0.5%), and acidic Asp (GAC, 0.3% but except *nad3*) were highly utilized across all PCGs. Among each codon family, the T-ending (49.9%) codons are used most frequently than codons ending in any of the three other nucleotides, with C-ending (2.7%) codons least used. Similarly, for both Haemonchidae and Trichostrongylidae the average nucleotide content of their mt PCGs are 26.0% (A) and 43.7% (T), compared to 25.4% (A), and 44.2% (T), respectively. In particular the *nad*4L was the most T-rich at 53.5 and 54.9%, in each family, respectively. Also the most commonly codon observed was TAA as the termination codon with the majority of initiation codons also being AT-rich.

The observed nucleotide biases also have substantial effects on codon usage patterns for both *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP mt genomes (**Table 2**). The most frequently used codons among the PCGs were: TGT (Cys, $N =$ 122 and 101 times used, 3.73% and 2.94% of the total), TTT (Phe, $N =$ 435 and 428, 13.32% and 12.47%), ATT (Ile, $N =$ 201 and 215, 6.15% and 6.26%), TTA (Leu, $N =$ 199 and 271, 6.09% and 7.90%), TTT (Asn, $N =$ 179 and 123, 5.48% and 3.60%), GTT (Val, $N =$ 143 and 119, 4.40% and 3.50%), and TAT (Tyr, $N =$ 181 and 198, 5.54% and 5.77%), respectively. In addition, T-ending codons are far more frequent than codons ending in any of the other three nucleotides among each codon family. For example, the least used codons included CCA (Pro, $N =$ 3 and 14, 0.09% and 0.41%), GCG (Ala, $N =$ 3 and 7, 0.09% and 0.20%), along with the C-ending codons GAC (Asp, $N =$ 8 and 10, 0.25% and 0.29%) and CTC (Leu, $N =$ 3 and 10, 0.09% and 0.29%).

Our pairwise alignments and synteny analysis of the complete mt genome nucleotide sequences for the *Haemonchus* and *Teladorsagia* species and strains revealed a common and highly variable A+T-rich non-coding region between 5,500 and 6,100 bp in each mt genome (**Figure 2** and **Supplementary Figure 3**). The A+T content of this region is unusually high in both *Haemonchus* (89.0%) and *Teladorsagia* (88.2%) mt genomes, and is located between *trnA* and *trnP* (or between genes *nad*5 and *nad*6). Among *Haemonchus*, the A+T-rich region is ∼600 bp in *H. contortus* (89.7%) and 155 bp in *H. placei* (87.1%), whereas the

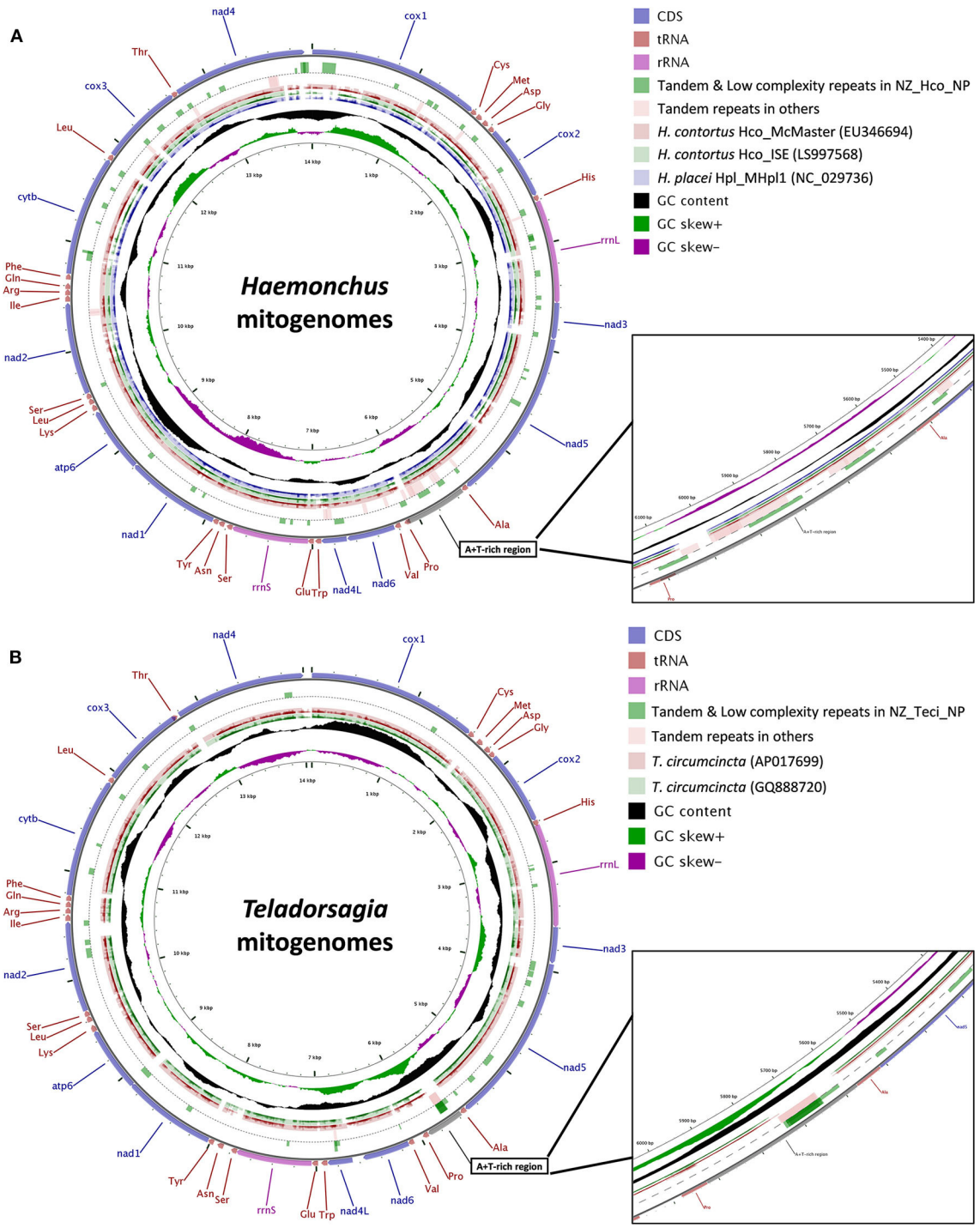


FIGURE 2 | Gene arrangement and distribution of *Haemonchus* **(A)** and *Teladorsagia* **(B)** mt genomes. **(A)** Gene map of the complete mitochondrial genomes for *H. contortus* Hco_McMaster (light red, EU346694), *H. contortus* Hco_ISE (light green, LS997568) and *H. placei* Hpl_MHpl1 (light purple, NC_029736). **(B)** Gene map of the complete mitochondrial genomes for *T. circumcincta* AP017699 (light red) and GQ888720 (light green). *H. contortus* NZ_Hco_NP **(A)** and *T. circumcincta* NZ_Teci_NP **(B)** were used as the reference sequences, respectively. The outermost ring indicates gene arrangement and distribution based on the primary sequence GenBank file. Tandem repeat regions are shown with green blocks (full length) for *H. contortus* NZ_Hco_NP (circle 2) and light red blocks for all other *Haemonchus* mt genomes (circle 3). Low complexity repeat regions for *H. contortus* NZ_Hco_NP are also shown with green blocks (half length) on circle 2. The next two **(A)** or three **(B)** rings represent BLAST hits obtained from BLASTX searches performed using a threshold of $1e^{-5}$ for protein sequences of genomes described above, in which the query sequences were translated in reading frames 1 and 2. The inner-most rings indicate the GC content information. The contents of the A+T-rich regions have been shown in more detail in expanded 10× zoomed views. nad1-6, NADH dehydrogenase subunits 1–6; cox1-3, cytochrome c oxidase subunits 1–3; atp6, ATPase subunit 6; cytb, cytochrome b.

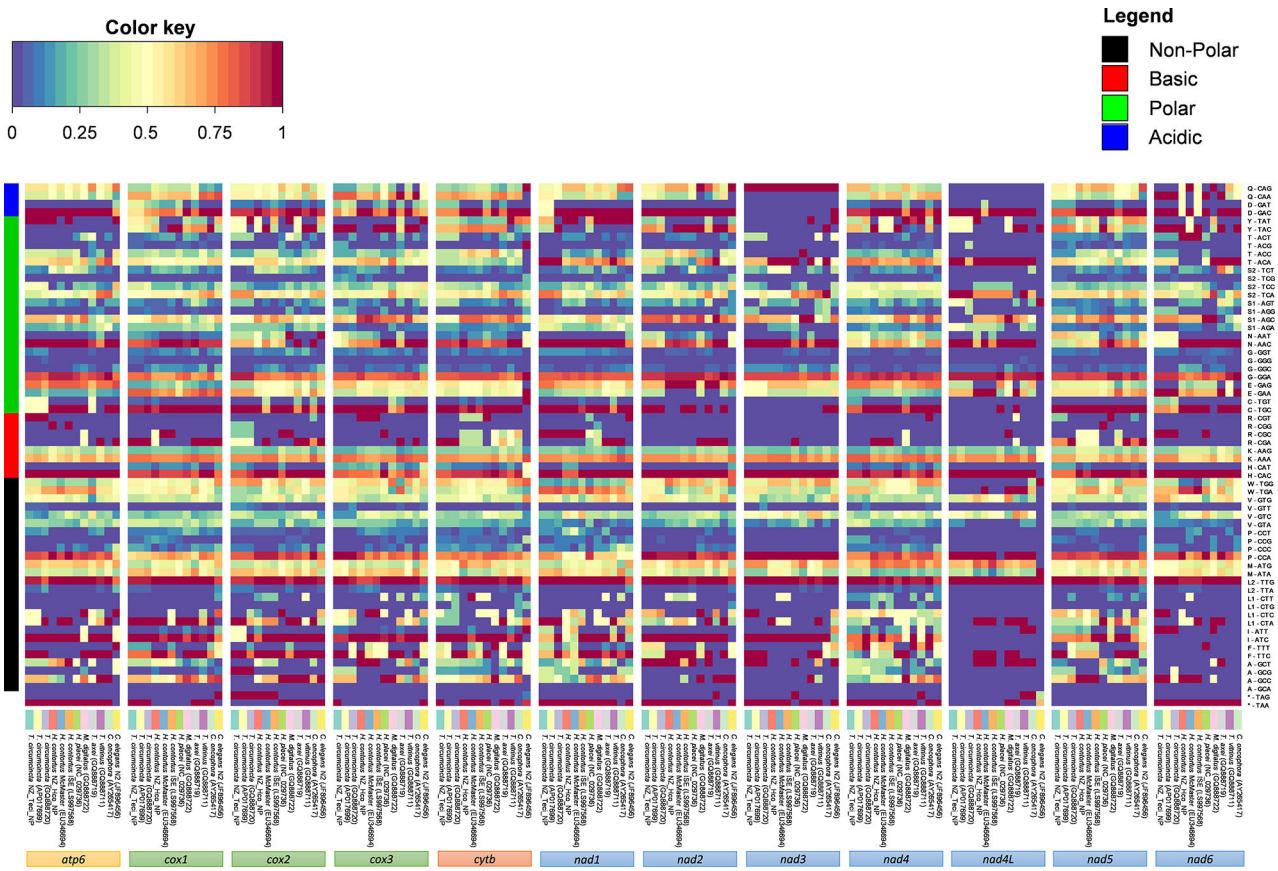


FIGURE 3 | Amino acid and codon usage of Trichostrongyloidea complete mt genomes. Heat map analysis was performed based on relative abundances of protein sequence content determined for all mitochondrial genes. The relative inferred codon abundances per mt gene and specific to each mt genome are shown using a heat color scheme (blue to red), indicating low to high relative abundance. Mt genomes are groped based on previous phylogenomic comparison of concatenated PCGs. *C. elegans* N2 (GenBank accession number JF896456) was also included for comparison. Amino acid families are colored to represent non-polar in black, basic in red, polar in green, and acidic in blue.

region is 340 bp in *Teladorsagia* (88.2%). A pairwise alignment of the nucleotide sequences of the A+T-rich regions of *H. contortus* NZ_Hco_NP (611 bp) and closely related *H. contortus* McMaster (EU346694, 578 bp) resulted in 91.5% identity. Interestingly, the region in *H. contortus* is among the largest among other nematodes studied to date with the exception of *As. suum* (886 bp) (Okimoto et al., 1992; Keddie et al., 1998; Hu et al., 2003).

Comparative analysis of the control (A+T-rich) region of *Haemonchus* mtDNAs identified three tandem repeat (TR) units (69, 120, and 62 bp) compared to only one (126 bp) found in *T. circumcincta* strains mt genomes (**Figure 2**). The number of repeat units within the A+T-rich region was variable depending on the mtDNAs. For example, the copy numbers of the A+T-rich region TRs for *T. circumcincta* NZ_Teci_NP corresponding to TTATAATTATTATATAATAATTA (23 bp), TATATATATATA AATTAATATAATTATTATTAATAAT (37 bp), and ATATTATA TATTATATATTA (20 bp) consensus units were 4, 2, and 4. In general, TRs of *Haemonchus* mt genomes were much more abundant and distributed randomly throughout the mtDNAs, whereas TRs were mostly in the intergenic regions of *T. circumcincta* mtDNA. Some TRs were found within the mitochondrial genes, such as those in *cox*1, *nad*5, *nad*6, and *nad*4, with also *nad*4L in *T. circumcincta* mtDNA. Overall, our analyses show that the number of tandem repeats was highly variable among different species and strains of *Haemonchus* and *Teladorsagia* and largely account for the variations among the mtDNA sequences lengths.

Based on our comparative genomics of A+T rich regions of complete *Haemonchus* species/strains and *T. circumcincta* strains mt genomes, we were able to differentiate individual strains with high confidence due to the high mutation rate and sequencing coverage for these regions. We propose that such regions can serve as ideal markers for future studies investigating nematode population structure (Jex et al., 2009; Gasser et al., 2012), population genetics (Blouin, 2002; Hu et al., 2004; Gasser et al., 2012), and phylogeny (Blouin et al., 1997; Blouin, 2002; Xu et al., 2015). For future work, we aim to design and test numerous primer pairs targeting the A+T-rich region between the *nad*5 and *nad*6 genes to develop accurate PCR-based molecular techniques to differentiate Trichostrongyloidea species

TABLE 2 | Codon usage analysis of PCGs in the mitochondrial genomes of *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP.

AA	Codon	N	/1,000[a]	Freq	AA	Codon	N	/1,000[a]	Freq
Ala	GCG	3|7	0.92|2.04	0.09|0.12	Pro	CCG	8|6	2.45|1.75	0.32|0.12
	GCA	12|10	3.67|2.91	0.38|0.18		CCA	3|14	0.92|4.08	0.12|0.27
	GCT	13|32	3.98|9.33	0.41|0.56		CCT	9|23	2.76|6.7	0.36|0.45
	GCC	4|8	1.22|2.33	0.13|0.14		CCC	5|8	1.53|2.33	0.2|0.16
Cys	TGT	122|101	37.35|29.44	0.84|0.84	Gln	CAG	13|18	3.98|5.25	0.3|0.37
	TGC	24|19	7.35|5.54	0.16|0.16		CAA	31|31	9.49|9.04	0.7|0.63
Asp	GAT	56|61	17.15|17.78	0.88|0.86	Arg	CGG	14|8	4.29|2.33	0.12|0.1
	GAC	8|10	2.45|2.91	0.13|0.14		CGA	6|1	1.84|0.29	0.05|0.01
Glu	GAG	17|32	5.21|9.33	0.3|0.43		CGT	15|21	4.59|6.12	0.12|0.26
	GAA	40|42	12.25|12.24	0.7|0.57		CGC	6|3	1.84|0.87	0.05|0.04
Phe	TTT	435|428	133.19|124.74	0.92|0.9		AGG	18|9	5.51|2.62	0.15|0.11
	TTC	36|46	11.02|13.41	0.08|0.1		AGA	62|38	18.98|11.08	0.51|0.47
Gly	GGG	25|43	7.65|12.53	0.26|0.3	Ser	AGT	69|70	21.13|20.4	0.41|0.43
	GGA	25|16	7.65|4.66	0.26|0.11		AGC	16|14	4.9|4.08	0.09|0.09
	GGT	38|72	11.64|20.99	0.4|0.5		TCG	20|10	6.12|2.91	0.12|0.06
	GGC	8|13	2.45|3.79	0.08|0.09		TCA	33|32	10.1|9.33	0.2|0.2
His	CAT	19|28	5.82|8.16	0.79|0.64		TCT	18|31	5.51|9.04	0.11|0.19
	CAC	5|16	1.53|4.66	0.21|0.36		TCC	13|6	3.98|1.75	0.08|0.04
Ile	ATA	133|103	40.72|30.02	0.37|0.31	Thr	ACG	11|15	3.37|4.37	0.15|0.17
	ATT	201|215	61.54|62.66	0.56|0.64		ACA	24|20	7.35|5.83	0.34|0.23
	ATC	27|19	8.27|5.54	0.07|0.06		ACT	24|48	7.35|13.99	0.34|0.55
Lys	AAG	49|37	15|10.78	0.3|0.27		ACC	12|5	3.67|1.46	0.17|0.06
	AAA	112|101	34.29|29.44	0.7|0.73	Val	GTG	47|42	14.39|12.24	0.18|0.17
Leu	TTG	128|160	39.19|46.63	0.32|0.29		GTA	59|62	18.06|18.07	0.22|0.26
	TTA	199|271	60.93|78.99	0.49|0.49		GTT	143|119	43.78|34.68	0.53|0.49
	CTG	11|17	3.37|4.95	0.03|0.03		GTC	19|18	5.82|5.25	0.07|0.07
	CTA	31|34	9.49|9.91	0.08|0.06	Tyr	TAT	181|198	55.42|57.71	0.84|0.82
	CTT	33|59	10.1|17.2	0.08|0.11		TAC	35|44	10.72|12.82	0.16|0.18
	CTC	3|10	0.92|2.91	0.01|0.02	Trp	TGG	44|48	13.47|13.99	1|1
Asn	AAT	179|123	54.81|35.85	0.85|0.9	Stop	TGA	32|44	9.8|12.82	0.14|0.16
	AAC	31|13	9.49|3.79	0.15|0.1		TAG	55|69	16.84|20.11	0.24|0.26
Met	ATG	52|56	15.92|16.32	1|1		TAA	142|154	43.48|44.88	0.62|0.58

H. contortus NZ_Hco_NP|T. circumcincta NZ_Teci_NP. [a]*Represents number of codons per 1,000 codons.*

members. Such molecular tools would enable easy species- and possibly strain-level identification with sufficient resolution and confidence from fecal samples that would also help prevent animal sacrifice for such studies.

Mitochondrial Phylogenomics

The super-family Trichostrongyloidea includes genera such as *Ostertagia*, *Teladorsagia*, *Trichostrongylus*, *Haemonchus*, *Cooperia*, *Nematodirus*, *Dictyocaulus*. To determine the phylogenetic relationship of the *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP with other *H. contortus* and *T. circumcincta* strains and members of the Trichostrongyloidea nematodes, the concatenated amino acid sequences predicted from 12 mtDNA PCGs were analyzed using BI, ML, and MP methods. Topologies of all trees inferred by the three different distance models and methods were identical (**Figure 4**), with phylogenetic relationships among the different Trichostrongyloidea species well-resolved with very high nodal support throughout. The presented phylogenomic tree of Trichostrongyloidea mt genomes, is consistent in topology with other published data (Chilton et al., 2006; Jex et al., 2009), and further support the hypothesis that the primary driver of early divergence is actually the site of infection in the host. Overall, rur results corroborate this hypothesis given that each of different species and strains of the Haemonchidae compared with Cooperidae and Trichostrongylidae families, are genetically closer to abomasal than to non-abomasal (intestine) Trichostrongyloidea, respectively. The anomaly to this hypothesis was *T. axei* that occupies the abomasum as the site of infection oppose to the small intestine of *T. vitrinus*, which implies the presence of other and yet uncharacterized drivers of genetic species diversity of these closely related parasites.

The phylogenomic tree based on the concatenated PCGs of the complete mt genomes (**Figure 4**), placed *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP in their respective clades with other *H. contortus* and *T. circumcincta* strains, which are in agreement with previous studies (Jacquiet et al., 1998; Nunes et al., 2013; Rohart et al., 2017; Palevich et al., 2018). However, our tree is not directly comparable to

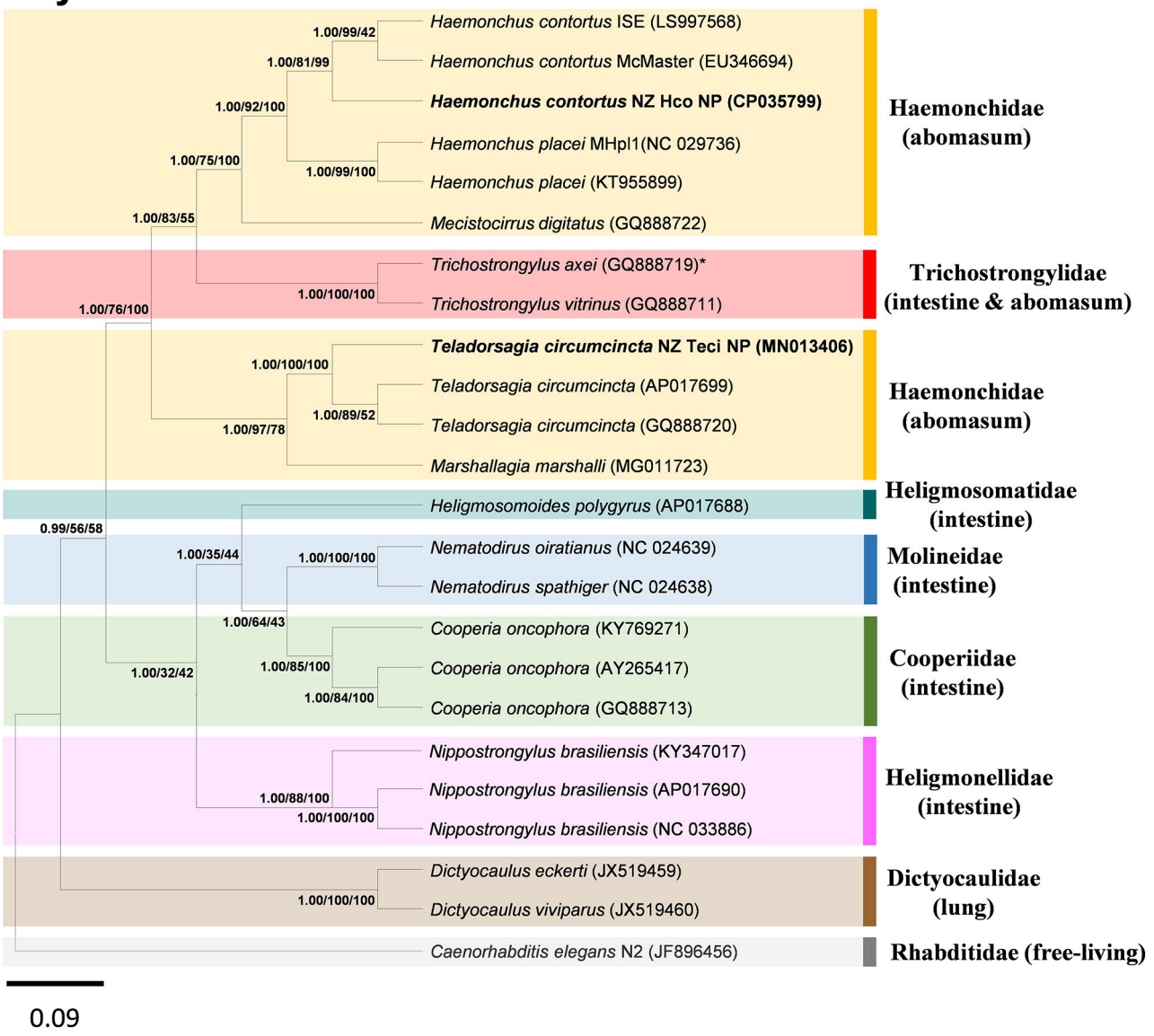


FIGURE 4 | Inferred phylogenetic relationship among the nematode superfamily Trichostrongyloidea. Phylogenetic analysis of the complete mitochondrial genomes of different nematode species or conspecific strains is based on the concatenated amino acid sequences of 12 protein-coding genes by Bayesian inference (BI), maximum likelihood (ML), and maximum parsimony (MP) analyses. *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP are in bold, with *C. elegans* N2 (GenBank accession number JF896456) used as the outgroup. *species with different site of infection to other members of family. The numbers along branches indicate posterior probabilities and bootstrap values resulting from different analyses in the order: BI/ML/MP. Members of different families were shown in colored boxes with the associated main site of infection in parenthesis. GenBank accession numbers are provided (in parentheses) for all reference sequences. Scale bar represents Posterior Probability.

previously described phylogenetic trees that performed different analyses and compared a broader set of taxa, but mainly because the trees were based on the ITS rDNA nuclear genetic markers (Chilton et al., 2001; Hoberg et al., 2004; Jex et al., 2008). While both *Haemoncus* and *Teladorsagia* strains grouped together in their respective families, *T. circumcincta* NZ_Teci_NP was clearly separated from the other strains. Interestingly, *H. contortus* NZ_Hco_NP and McMaster (Australia) strains grouped together and originate from nearby geographic locations, but separate from *H. contortus* ISE (East Africa). This observation corroborates our hypothesis on the impact of environmental constraints and preferences acting as secondary drivers of within-species diversity. Overall, we are in agreement with previous phylogenetic studies suggesting that ITS rDNA nuclear genetic markers alone may not provide sufficient information to reveal higher taxonomic level relationships within the Trichostrongyloidea. Our findings indicate that a three-pronged approach that incorporates phylogenetic inertia,

pangenome structure/features and environmental data in order to understand the mitochondrial genome evolution.

CONCLUSIONS

This study compares the recently sequenced complete mitochondrial genomes of sibling species of parasitic roundworms, *H. contortus* and *T. circumcincta*, two of the most economically important and common pathogenic nematodes infecting small ruminants worldwide. We explored the mitochondrial pangenome features and phylogenomic relationships to assess species- and strain-level diversity among Trichostrongyloidea nematodes and with available mt genome in public databases. Our analyses corroborate previous studies showing that our *H. contortus* NZ_Hco_NP and *T. circumcincta* NZ_Teci_NP strains position in their respective clades. Future work should focus on utilizing our insights on the highly variable regions of *Haemoncus* and *Teladorsagia* conspecific strains to develop cost-effective DNA-based approaches for novel parasite management and control strategies. The complete mt genomes of the New Zealand *H. contortus* and *T. circumcincta* field strains are important contributions to our understanding of meta-population connectivity as well as species- and strain-level evolution in nematodes. With the continuing improvements in sequencing technology combined with a community effort we may be able to reconstruct the true origins of *Haemoncus* and *Teladorsagia*, as well as other parasitic nematodes that are of a global interest.

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

NP conceived and designed the project, analyzed and interpreted the data, and wrote the manuscript. PM provided bioinformatics support for the project. MM and Y-JC provided resources and critically revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was supported by the Agricultural and Marketing Research and Development Trust (AGMARDT) Postdoctoral Fellowship Programme, grant number P17001.

ACKNOWLEDGMENTS

Special thanks to PM, AgResearch Limited, Grasslands Research Centre, Palmerston North, New Zealand, for his contribution to data analysis. This manuscript has been released as a pre-print at Research Square (Palevich et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.573395/full#supplementary-material

Supplementary Figure 1 | Predicted secondary structures of the 22 tRNAs in the *H. contortus* NZ_Hco_NP (left) and *T. circumcincta* NZ_Teci_NP (right) mitochondrial genomes. The two-dimensional predicted layout predicted RNA secondary structures is shown with canonical base-pair (BP) in blue, and non-canonical, lone-pairs and triplets in green.

Supplementary Figure 2 | Predicted secondary structures of the *rrn*S **(A)** and *rrn*L **(B)** genes in the *H. contortus* NZ_Hco_NP (left) and *T. circumcincta* NZ_Teci_NP (right) mitochondrial genomes. The two-dimensional predicted layout predicted RNA secondary structures were predicted based on the energy model of Mathews et al. (2004). The minimum free energy (MFE) structure of hairpins is colored according to the base-pairing probabilities (red, high; green, mid; blue, low). Blue and red circles around nucleotides represent the beginning and the end of molecules, respectively.

Supplementary Figure 3 | Dot plots showing *Haemonchus* and *Teladorsagia* species and strain-level synteny. The horizontal and vertical axes represent the entire translated mt genome nucleotide sequences, respectively. Each aligning gap-free segment with more than 50% identity is plotted as a black line or dot. Analysis was performed using chaining coverage with PipMaker (Schwartz et al., 2000).

REFERENCES

Adachi, J., and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42, 459–468. doi: 10.1007/BF02498640

Amarante, A., Bagnola, J. Jr., Amarante, M., and Barbosa, M. (1997). Host specificity of sheep and cattle nematodes in Saõ Paulo state, Brazil. *Vet. Parasitol.* 73, 89–104. doi: 10.1016/S0304-4017(97)-8

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Cambridge: Babraham Bioinformatics, Babraham Institute.

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., et al. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature* 392, 71–75. doi: 10.1038/32160

Blouin, M. S. (2002). Molecular prospecting for cryptic species of nematodes: mitochondrial DNA versus internal transcribed spacer. *Int. J. Parasitol.* 32, 527–531. doi: 10.1016/S0020-7519(01)00357-5

Blouin, M. S., Yowell, C. A., Courtney, C. H., and Dame, J. B. (1997). *Haemonchus placei* and *Haemonchus contortus* are distinct species based on mtDNA evidence. *Int. J. Parasitol.* 27, 1383–1387. doi: 10.1016/S0020-7519(97)00125-2

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

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucl. Acids Res.* 27, 1767–1780. doi: 10.1093/nar/27.8.1767

Chilton, N. B., Huby-Chilton, F., Gasser, R. B., and Beveridge, I. (2006). The evolutionary origins of nematodes within the order Strongylida are related to predilection sites within hosts. *Mol. Phylogenet. Evol.* 40, 118–128. doi: 10.1016/j.ympev.2006.01.003

Chilton, N. B., Newton, L. A., Beveridge, I., and Gasser, R. B. (2001). Evolutionary relationships of trichostrongyloid nematodes (Strongylida) inferred from ribosomal DNA sequence data. *Mol. Phylogenet. Evol.* 19, 367–386. doi: 10.1006/mpev.2001.0938

Choi, Y.-J., Bisset, S. A., Doyle, S. R., Hallsworth-Pepin, K., Martin, J., Grant, W. N., et al. (2017). Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLoS Genet*. 13:e1006857. doi: 10.1371/journal.pgen.1006857

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25:1974. doi: 10.1093/bioinformatics/btp250

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucl. Acids Res.* 45:e18. doi: 10.1093/nar/gkw955

dos Santos, L. L., Prosdocimi, F., Lima, N. C. B., da Costa, I. R., Cardoso, D. C., Drummond, M. G., et al. (2017). Comparative genomics and phylogenomics of Trichostrongyloidea mitochondria reveal insights for molecular diagnosis and evolutionary biology of nematode worms. *Gene Rep.* 9, 65–73. doi: 10.1016/j.genrep.2017.09.002

Gasser, R., Schwarz, E., Korhonen, P., and Young, N. (2016). "Understanding *Haemonchus contortus* better through genomics and transcriptomics," in *Advances in Parasitology*, eds R. B. Gasser and G. V. Samson-Himmelstjerna (Melbourne; Berlin: Elsevier), 519–567.

Gasser, R. B., Bott, N. J., Chilton, N. B., Hunt, P., and Beveridge, I. (2008). Toward practical, DNA-based diagnostic methods for parasitic nematodes of livestock-bionomic and biotechnological implications. *Biotechnol. Adv.* 26, 325–334. doi: 10.1016/j.biotechadv.2008.03.003

Gasser, R. B., Hu, M., Chilton, N. B., Campbell, B. E., Jex, A. J., Otranto, D., et al. (2006). Single-strand conformation polymorphism (SSCP) for the analysis of genetic variation. *Nat. Protocols* 1:3121. doi: 10.1038/nprot.2006.485

Gasser, R. B., Jabbar, A., Mohandas, N., Höglund, J., Hall, R. S., Littlewood, D. T. J., et al. (2012). Assessment of the genetic relationship between Dictyocaulus species from *Bos taurus* and *Cervus elaphus* using complete mitochondrial genomic datasets. *Parasites Vectors* 5:241. doi: 10.1186/1756-3305-5-241

Grant, J. R., Arantes, A. S., and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics* 13:202. doi: 10.1186/1471-2164-13-202

Gutell, R. R., Gray, M. W., and Schnare, M. N. (1993). A compilation of large subunit (23S and 23S-like) ribosomal RNA structures. *Nucl. Acids Res.* 21, 3055–3074. doi: 10.1093/nar/21.13.3055

Hoberg, E. P., Lichtenfels, J. R., and Gibbons, L. (2004). Phylogeny for species of Haemonchus (Nematoda: Trichostrongyloidea): considerations of their evolutionary history and global biogeography among Camelidae and Pecora (Artiodactyla). *J. Parasitol.* 90, 1085–1102. doi: 10.1645/GE-3309

Hu, M., Chilton, N. B., El-Osta, Y. G. A., and Gasser, R. B. (2003). Comparative analysis of mitochondrial genome data for *Necator americanus* from two endemic regions reveals substantial genetic variation. *International Journal for Parasitology*. 33, 955–963. doi: 10.1016/S0020-7519(03)00129-2

Hu, M., Chilton, N. B., and Gasser, R. B. (2004). The mitochondrial genomics of parasitic nematodes of socio-economic importance: recent progress, and implications for population genetics and systematics. *Adv. Parasitol.* 56, 134–213. doi: 10.1016/S0065-308X(03)56003-1

Hu, M., and Gasser, R. B. (2006). Mitochondrial genomes of parasitic nematodes–progress and perspectives. *Trends Parasitol.* 22, 78–84. doi: 10.1016/j.pt.2005.12.003

Hu, M., Jex, A. R., Campbell, B. E., and Gasser, R. B. (2007). Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nat. Protocols* 2:2339. doi: 10.1038/nprot.2007.358

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Jacquiet, P., Cabaret, J., Thiam, E., and Cheikh, D. (1998). Host range and the maintenance of *Haemonchus* spp. in an adverse arid climate. *Int. J. Parasitol.* 28, 253–261. doi: 10.1016/S0020-7519(97)00185-9

Jex, A. R., Hall, R. S., Littlewood, D. T. J., and Gasser, R. B. (2009). An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucl. Acids Res.* 38, 522–533. doi: 10.1093/nar/gkp883

Jex, A. R., Hu, M., Littlewood, D. T. J., Waeschenbach, A., and Gasser, R. B. (2008). Using 454 technology for long-PCR based sequencing of the complete mitochondrial genome from single *Haemonchus contortus* (Nematoda). *BMC Genomics* 9:11. doi: 10.1186/1471-2164-9-11

Kaplan, R. M. (2004). Drug resistance in nematodes of veterinary importance: a status report. *Trends Parasitol.* 20, 477–481. doi: 10.1016/j.pt.2004.08.001

Kaplan, R. M., and Vidyashankar, A. N. (2012). An inconvenient truth: global worming and anthelmintic resistance. *Vet. Parasitol.* 186, 70–78. doi: 10.1016/j.vetpar.2011.11.048

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Keddie, E. M., Higazi, T., and Unnasch, T. R. (1998). The mitochondrial genome of *Onchocerca volvulus*: sequence, structure and phylogenetic analysis. *Mol. Biochem. Parasitol.* 95, 111–127. doi: 10.1016/S0166-6851(98)00102-9

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

Lavrov, D. V., and Brown, W. M. (2001). *Trichinella spiralis* mtDNA: a nematode mitochondrial genome that encodes a putative ATP8 and normally structured tRNAs and has a gene arrangement relatable to those of coelomate metazoans. *Genetics* 157, 621–637. Available online at: https://www.genetics.org/content/157/2/621

Le, V. S., Dang, C. C., and Le, Q. S. (2017). Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. *BMC Evol Biol.* 17:136. doi: 10.1186/s12862-017-0987-y

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol.* 6:26. doi: 10.1186/1748-7188-6-26

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7287–7292. doi: 10.1073/pnas.0401799101

Morgulis, A., Gertz, E. M., Schäffer, A. A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* 13, 1028–1040. doi: 10.1089/cmb.2006.13.1028

Nunes, R. L., dos Santos, L. L., Bastianetto, E., de Oliveira, D. A. A., and Brasil, B. S. A. F. (2013). Frequency of benzimidazole resistance in *Haemonchus contortus* populations isolated from buffalo, goat and sheep herds. *Rev. Brasil. Parasitol. Vet.* 22, 548–553. doi: 10.1590/S1984-29612013000400015

Okimoto, R., Macfarlane, J., Clary, D., and Wolstenholme, D. (1992). The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics* 130, 471–498.

Palevich, N., Britton, C., Kamenetzky, L., Mitreva, M., de Moraes Mourão, M., Bennuru, S., et al. (2018). Tackling hypotheticals in helminth genomes. *Trends Parasitol.* 34, 179–183. doi: 10.1016/j.pt.2017.11.007

Palevich, N., Maclean, P., Baten, A., Scott, R., and Leathwick, D. M. (2019a). The complete mitochondrial genome of the New Zealand parasitic roundworm *Haemonchus contortus* (Trichostrongyloidea: Haemonchidae) field strain NZ_Hco_NP. *Mitochondrial DNA Part B* 4, 2208–2210. doi: 10.1080/23802359.2019.1624634

Palevich, N., Maclean, P. H., Baten, A., Scott, R. W., and Leathwick, D. M. (2019c). The genome sequence of the anthelmintic-susceptible New Zealand *Haemonchus contortus*. *Genome Biol. Evol.* 11, 1965–1970. doi: 10.1093/gbe/evz141

Palevich, N., Maclean, P. H., Choi, Y. J., and Mitreva, M. (2020). Characterization of the complete mitochondrial genomes of two sibling species of parasitic

roundworms, *Haemonchus contortus* and *Teladorsagia circumcincta*. *Res. Square*. 1, 1–37. doi: 10.21203/rs.3.rs-24858/v1

Palevich, N., Maclean, P. H., Mitreva, M., Scott, R., and Leathwick, D. (2019b). The complete mitochondrial genome of the New Zealand parasitic roundworm *Teladorsagia circumcincta* (Trichostrongyloidea: Haemonchidae) field strain NZ_Teci_NP. *Mitochondrial DNA Part B* 4, 2869–2871. doi: 10.1080/23802359.2019.1660241

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Bellingham, WA.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol*. 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., and Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238, 195–209. doi: 10.1016/S0378-1119(99)00270-X

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res*. 10, 577–586. doi: 10.1101/gr.10.4.577

Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* 10, 1–13. doi: 10.1038/s41467-019-13395-9

Stevenson, L. A., Chilton, N. B., and Gasser, R. B. (1995). Differentiation of *Haemonchus placei* from, *H. contortus* (Nematoda: Trichostrongylidae) by the ribosomal DNA second internal transcribed spacer. *Int. J. Parasitol.* 25, 483–488. doi: 10.1016/0020-7519(94)00156-I

Sutherland, I., and Scott, I. (2010). *Gastrointestinal Nematodes of Sheep and Cattle: Biology and Control*. Palmerston North: Wiley-Blackwell.

Swofford, D. L. (2001). *Paup*: Phylogenetic Analysis Using Parsimony (and Other methods) 4.0. B5*. Sunderland.

Tang, Y. T., Gao, X., Rosa, B. A., Abubucker, S., Hallsworth-Pepin, K., Martin, J., et al. (2014). Genome of the human hookworm *Necator americanus*. *Nat. Genet*. 46:261. doi: 10.1038/ng.2875

Veglia, F. (1915). *The Anatomy and Life-History of Haemonchus contortus (Rud.)*. Pretoria: Government Printer and Stationery Office.

Vlassoff, A., Leathwick, D., and Heath, A. (2001). The epidemiology of nematode infections of sheep. *New Zealand Vet. J.* 49, 213–221. doi: 10.1080/00480169.2001.36235

Wit, J., and Gilleard, J. S. (2017). Resequencing helminth genomes for population and genetic studies. *Trends Parasitol*. 33, 388–399. doi: 10.1016/j.pt.2017.01.009

Wolstenholme, D. R. (1992). "Animal mitochondrial DNA structure and evolution," in *International Review of Cytology*, eds D. R. Wolstenholme and K. W. Jeon (Salt Lake City, UT; Knoxville: Elsevier), 173–216.

Wolstenholme, D. R., Macfarlane, J. L., Okimoto, R., Clary, D. O., and Wahleithner, J. A. (1987). Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. *Proc. Natl. Acad. Sci. U.S.A.* 84, 1324–1328. doi: 10.1073/pnas.84.5.1324

Wolstenholme, D. R., Okimoto, R., and Macfarlane, J. L. (1994). Nucleotide correlations that suggest tertiary interactions in the TV-replacement loop-containing mitochondrial tRNAs of the nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Nucl. Acids Res*. 22, 4300–4306. doi: 10.1093/nar/22.20.4300

Xu, W-W, Qiu, J-H, Liu, G-H, Zhang, Y., Liu, Z-X, Duan, H. et al. (2015). The complete mitochondrial genome of *Strongylus equinus* (Chromadorea: Strongylidae): comparison with other closely related species and phylogenetic analyses. *Exp Parasitol*. 159, 94–99. doi: 10.1016/j.exppara.2015.08.012

Yokobori, S-, and i, Pääbo, S. (1995). tRNA editing in metazoans. *Nature* 377:490. doi: 10.1038/377490a0

Zarowiecki, M., Huyse, T., and Littlewood, D. (2007). Making the most of mitochondrial genomes-markers for phylogeny, molecular ecology and barcodes in Schistosoma (Platyhelminthes: Digenea). *Int. J. Parasitol*. 37, 1401–1418. doi: 10.1016/j.ijpara.2007.04.014

Conflict of Interest: NP was employed by the company AgResearch Limited.

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

Copyright © 2020 Palevich, Maclean, Choi and Mitreva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 October 2020
doi: 10.3389/fmicb.2020.595902

Molecular Pap Smear: Validation of HPV Genotype and Host Methylation Profiles of *ADCY8*, *CDH8*, and *ZNF582* as a Predictor of Cervical Cytopathology

Jane Shen-Gunther[1,2], Qingqing Xia[3], Winfred Stacey[3] and Heisy B. Asusta[4]*

[1] Gynecologic Oncology & Clinical Investigation, Department of Clinical Investigation, Brooke Army Medical Center, Fort Sam Houston, TX, United States, [2] Department of Molecular Medicine, Cancer Therapy and Research Center, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States, [3] Department of Clinical Investigation, Brooke Army Medical Center, Fort Sam Houston, TX, United States, [4] Department of Obstetrics and Gynecology, Brooke Army Medical Center, Fort Sam Houston, TX, United States

OPEN ACCESS

Edited by:
Zhan Zhou,
Zhejiang University, China

Reviewed by:
Timothy Lilburn,
Novozymes (United States), United States
Xiang Gao,
Loyola University Chicago, United States

***Correspondence:**
Jane Shen-Gunther
jane.shengunther.mil@mail.mil;
shengunther@livemail.uthscsa.edu

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

Received: *17 August 2020*
Accepted: *22 September 2020*
Published: *15 October 2020*

Citation:
Shen-Gunther J, Xia Q, Stacey W and Asusta HB (2020) Molecular Pap Smear: Validation of HPV Genotype and Host Methylation Profiles of ADCY8, CDH8, and ZNF582 as a Predictor of Cervical Cytopathology. Front. Microbiol. 11:595902. doi: 10.3389/fmicb.2020.595902

Primary high-risk Human Papillomavirus (hrHPV) screening has recently become an accepted standalone or co-test with conventional cytology. Unfortunately, hrHPV singularly lacks specificity for cytopathological grade. However, mechanisms and markers of evolving virus-host interactions at the epigenome level may be harnessed as a better predictor of carcinogenesis. This study aimed to validate and expand the clinical performance of a multiparametric biomarker panel, referred to as the "Molecular Pap smear" based, on HPV genotype and *ADCY8*, *CDH8* and *ZNF582* CpG-methylation as a predictive classifier of cervical cytology. This prospective, cross-sectional study used an independent cohort of residual liquid-based cytology for HPV genotyping and epigenetic analysis. Extracted DNA underwent parallel PCR using 3 primer sets for HPV DNA amplification. HPV-infected samples were genotyped by Sanger sequencing. Promoter methylation levels of 3 tumor suppressor genes were quantified by bisulfite-pyrosequencing of genomic DNA on the newest high-resolution PyroMark Q48 platform. Logistic model performance was compared, and model parameters were used to predict and classify binary cytological outcomes. A total of 883 samples were analyzed. HPV DNA positivity correlated with worsening grade: 125/237 (53%) NILM; 136/235 (58%) ASCUS; 222/229 (97%) LSIL; and 157/182 (86%) HSIL samples. The proportion of carcinogenic HPV-types in PCR-positive sequenceable samples correlated with worsening grade: NILM 34/98 (35%); ASCUS 50/113 (44%); LSIL 92/214 (43%); HSIL 129/152 (85%). Additionally, *ADCY8*, *CDH8*, and *ZNF582* methylation levels increased in direct correlation with worsening grade. Overall, the multi-marker modeling parameters predicted binarized cytological outcomes better than HPV-type alone with significantly higher area under the receiver operator curve (AUC)s, respectively: NILM vs. > NILM (AUC 0.728 vs. 0.709); NILM/ASCUS vs. LSIL/HSIL (AUC 0.805 vs. 0.776); and <HSIL vs. HSIL (AUC 0.830 vs. 0.761). Our expanded findings

validated the multivariable prediction model developed for cytological classification. The sequencing-based "Molecular Pap smear" outperformed HPV-type alone in predicting four grades of cervical cytology. Additional host epigenetic markers that evolved with disease progression decidedly contributed to the overall classification accuracy.

Keywords: carcinogenesis, DNA methylation, epigenetic modification, evolution, host-pathogen interactions, human papillomavirus infection, pap smear, pyrosequencing

INTRODUCTION

May 13, 2019 marked the 136th birthday of Dr. George Papanicolaou who was honored globally with an iconic image of his pioneering work on cervical cancer screening and the Papanicolaou (Pap) smear (Papanicolaou and Traut, 1941; Maxouris, 2019). Armed with a microscope, glass slides, cellular stains, and his wife, Andromachi (his life-long control subject), Papanicolaou changed the world of cancer prevention and founded a new scientific discipline called cytopathology (Carmichael and Cameron, 1973). Although 80 years have passed since the invention of the Pap smear, impenetrable economic and social barriers have prevented this life-saving test from reaching impoverished regions of the world (Canfell et al., 2020).

On World Cancer Day (4 February 2020), the World Health Organization (WHO) announced that a 60% increase in cancer cases worldwide is projected over the next two decades (World Health Organization [WHO] News Release, 2020). Currently, a staggering 18 million new cancer cases are diagnosed globally each year of which 13% (2.2 million) are caused by infectious agents (Wild et al., 2020). Human papillomavirus (HPV) ranks second only to Helicobacter pylori as the primary infectious cause of cancer. Annually, HPV is responsible for 570,000 new cervical and 120,000 other anogenital and oropharyngeal cancer cases (Wild et al., 2020). Low- and middle-income countries continue to carry the highest cancer burden and will incur the greatest increase in cancer incidence and mortality in the years to come due to inadequate resources for cancer prevention and early detection (World Health Organization [WHO], 2014, 2020).

To circumvent the formidable economic and infrastructural requirements associated with cytology-based screening programs and limitations of commercial HPV diagnostics, we developed a molecular diagnostic test called "Molecular Pap smear" which is based on HPV genotyping and quantitative DNA methylation. Fundamentally, the test harnesses the evolutionary characteristics of the pathogen, the host and infected host-tissue throughout carcinogenesis for use as biomarkers (Bosch et al., 2013; Chen et al., 2018). Our prior investigation had shown a loss of HPV genotypic diversity and gain of clonal dominance by carcinogenic genotypes in high-grade versus low-grade cytology (Shen-Gunther et al., 2017). Additionally, distinct patterns of loci-specific promoter hypermethylation were discovered and were consistent with the underlying mechanism of HPV E6 and E7 oncoprotein induced DNA methyltransferase activity and ensuing gene silencing (Durzynska et al., 2017). Our initial study which analyzed ∼300 cervical cytology samples indicated HPV genotype and host promoter methylation may perform well as a molecular classifier of cervical cytopathology. Furthermore, this work showed that the positive correlation between 1) HPV carcinogenicity, 2) *ADCY8, CDH8*, and *ZNF582* promoter hypermethylation as well as 3) grade of cervical pathology were quantifiable and distinctive.

In this study, we aimed to validate and expand the clinical performance of our multiparametric biomarker panel with an independent sample set inclusive of four cytological categories [negative for intraepithelial lesion or malignancy (NILM), atypical squamous cells of undetermined significance (ASC-US), low-grade squamous intraepithelial lesion (LSIL), and high-grade squamous intraepithelial lesion (HSIL)]. Secondarily, we sought to validate quantitative CpG-methylation by pyrosequencing (PSQ) on the newest high-resolution PyroMark 48 Autoprep platform, which to our knowledge has not yet been reported in the literature in contrast to the preceding models, Q24 and Q96 (Johannessen et al., 2018). The results of this study will aid in the translation of our current discoveries based on virus-host evolutionary characteristics of HPV-induced carcinogenesis into a screening test that is more accurate, affordable, and widely available to improve global health (Turajlic et al., 2015; Beerenwinkel et al., 2016).

MATERIALS AND METHODS

Subjects and Samples

This prospective cross-sectional study was conducted after approval by the Institutional Review Board of Brooke Army Medical Center (BAMC), Texas. Cervical specimens were collected from adult women ≥ 18 years of age undergoing cervical cytology screening. Cervical specimens with low cellularity (cell pellet volume < 200 uL) were excluded from analysis. Liquid-based cytology collected for clinical testing at the Department of Pathology of BAMC was consecutively procured after completion of analysis for cytological diagnosis. Samples were stored at room temperature until weekly batch DNA extraction. Demographic data were abstracted from the electronic health record (AHLTA) of the Department of Defense (DoD) and

Abbreviations: *ADCY8*, adenylate cyclase 8; ASC-US, atypical squamous cells of undetermined significance; AUC, Area under the receiver operator curve; CARC, carcinogenic HPV; *CDH8,* cadherin 8, type 2; CIN, cervical intraepithelial neoplasia; chr, chromosome; gDNA, genomic DNA; HPV, Human Papillomavirus; HSIL, high-grade squamous intraepithelial lesion; IARC, International Agency for Research on Cancer; LSIL, low-grade squamous intraepithelial lesion; NA, not available/identifiable by BLAST; NILM, negative for intraepithelial lesion or malignancy; NOT CARC, not carcinogenic; NS, not significant; Pap, Papanicolaou smear; POSS CARC, possibly carcinogenic; PSQ, pyrosequencing; Q48, PyroMark Q48 platform; Q96, PyroMark Q96 platform; ROC, receiver operating characteristic; SCC, squamous cell carcinoma; *ZNF582,* zinc finger protein 582.

code-linked to each specimen. Four categories of samples were collected until target accrual numbers totaling 883 samples were met [NILM ($n = 237$), ASCUS ($n = 235$), LSIL ($n = 229$), and HSIL ($n = 182$)].

Laboratory Schema

The laboratory schema is illustrated in **Figure 1A**. After sample collection, cellular DNA is extracted from cervical cytology. The DNA is subjected to HPV DNA amplification, sequencing, and genotyping. For DNA methylation analysis, the genomic DNA undergoes bisulfite conversion and PSQ. The results derived from HPV genotyping and methylation quantification are analyzed for correlation with the cytological grade. **Figure 1B** shows representative images of the 4 categories of cervical cytology used in this study and binarization schema of cytological outcomes for logistic regression, prediction, and classification. Morphological features and differences among the cytological categories are highlighted by the relative size and distribution of organelles. The PyroMark Q48 instrument and PSQ assays are shown in **Figure 1C**.

HPV DNA Amplification

Cellular DNA extraction and HPV DNA amplification were performed as described previously (Shen-Gunther et al., 2016). Briefly, individual liquid-based cytology samples (10 mL) were centrifuged to obtain the cell pellet (200–250 μL) for DNA extraction using the QIAamp DNA Mini kit in a QIAcube robotic workstation (Qiagen). The purified DNA in 150 μL of eluent was quantified by spectrophotometry using the QIAxpert (Qiagen) and stored at -20°C prior to amplification. For HPV DNA amplification, 3 consensus primer sets: (1) MY09/11, (2) FAP59/64, and (3) GP-E6-3F/GP-E7-5B/GP-E7-6B were used to amplify 2 distinct regions of the HPV L1 and E6/E7 genes for genotype identification, respectively, as illustrated in **Supplementary Figure 1** (Resnick et al., 1990; Forslund et al., 1999; Sotlar et al., 2004; Shen-Gunther and Yu, 2011). Functionally, E6 and E7 codes for oncoproteins which inactivate two respective host cellular proteins p53 and RB leading to malignant transformation. L1 codes for the structural capsid protein which is essential for viral binding and entry into host tissues (Buck et al., 2013). AmpliTaq Gold 360 Master Mix (Life Technologies) and Qiagen Multiplex PCR Plus kit (Qiagen) were used with the doublet and triplet primer sets, respectively. The PCR cycling protocols for the 3 primer sets were also performed as described previously (Shen-Gunther et al., 2016). After amplification, high-resolution capillary gel electrophoresis using the QIAxcel (Qiagen) was performed to detect HPV DNA amplicons for follow-on DNA sequencing.

HPV DNA Sequencing, Genotyping, and Phylogenetic Analysis

Sanger sequencing of the amplicons (∼200 ng DNA/sample) was performed by using sequencing primers MY11, FAP59, and GP-E6-3F at Eurofins Genomics (Louisville, KY). Sequence quality was assessed using the Sequence Scanner 2.0 (appliedbiosystems.com), where a "high quality" Trace Score (TS) (average base call quality value as measured by phred quality score) was defined as ≥20 and a QV20 + value (total number of bases in the sequence with TS ≥ 20) as ≥100. Quality sequences were filter selected for entry into the Basic Local Alignment Search Tool (BLAST®) and queried against HPV sequences in GenBank® under Virus Taxonomy ID#151340 (Shen-Gunther and Yu, 2011) using the CLC Genomics Workbench v20.0.4 (Redwood City, CA). The HPV genotype was based on the most homologous and significant result with a minimum Expected Value (*E*-value) $< 10E^{-50}$. For each sample, if the identified HPV genotype differed among the primer-specific amplicons, the genotype with the lowest *E*-value took precedence. Furthermore, if the *E*-values were equal between sequenced amplicons (e.g., *E*-value = 0), the genotype was assigned in descending order of precedence by the primer-specific amplicon: E6/E7, MY09/11, and FAP59/64. The rationale for this ranking was based on the clinical significance of HPV E6/E7 over L1 gene function in respect to carcinogenic potential. The proportions of samples in which HPV was detected according to genotype and genotype-specific carcinogenic potential within each cytological category were compared.

To explore the evolutionary relationship of all HPV genotypes identified in the clinical samples, a representative phylogenetic tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987) after concatenating the aligned, E6 (∼477 bp), E7 (∼297 bp), and L1 (∼1,576 bp) reference coding sequences from Papillomavirus Episteme[1] by MUSCLE (Edgar, 2004). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004). Codon positions included were 1st + 2nd + 3rd + Non-coding. Positions containing gaps or missing data were eliminated. Bootstrap analysis using 1,000 replicates was performed to evaluate the reliability of the inferred tree (Felsenstein, 1985). Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018).

DNA Methylation Quantification

For DNA methylation profiling of cervical cytology, extracted genomic DNA (≥ 20 ng/uL) was bisulfite-converted to convert unmethylated cytosine residues to uracil using the EpiTect Fast 96 Bisulfite Conversion kit (Qiagen) per the manufacturer's instructions. Loci-specific PCR amplification of the bisulfite-converted gDNA (10–20 ng) was performed using 3 primer pairs targeting *ADCY8*, *CDH8*, or *ZNF582* combined with Pyromark PCR Master Mix (Qiagen) per the manufacturer's instructions. The PSQ assays including primer sequences were performed as described previously (Shen-Gunther et al., 2016). The PCR reaction (25 uL volume) and cycling protocol were performed per manufacturer's instructions and are described as follows: activation [95°C × 15 min]; 45 cycles of 3-step cycling [94°C × 30 s, 56°C × 30 s, 72°C × 30 s]; and final extension [72°C × 10 min].

The biotinylated PCR product was analyzed using high-resolution capillary gel electrophoresis (QIAxcel) for expected size in base pairs and adequacy of DNA concentration (>1 ng/uL)

[1] http://pave.niaid.nih.gov

prior to PSQ (**Supplementary Figure 2**). The PCR product (10 uL) and magnetic beads (3 uL) were pipetted into the PyroMark Q48 disk wells and loaded on the Q48 instrument (Qiagen) for PSQ (**Figure 1C**). Post-run results for CpG methylation quantification were analyzed using the Q48 Autoprep software 2.4.2 on CpG analysis mode and visualized as sequence-specific pyrograms (**Figure 1C**). The individual CpG-methylation levels (%) of each sample were joined with HPV status to construct the multivariable logistic models below.

Definitions, Variable Coding, and Logistic Modeling

The WHO International Agency for Research on Cancer (IARC) Working Group classifies HPV carcinogenic potential into three primary categories (International Agency for Research on Cancer [IARC], 2012) (1) carcinogenic (HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 68) (2) possibly carcinogenic (HPV types 26, 30, 34, 53, 66, 67, 69, 70, 73, 82, 85, and 97) and (3) not classifiable/probably not carcinogenic (HPV types 6, 11, and all others) (Schiffman et al., 2009; Bernard et al., 2010).

To compare the prevalence of HPV genotypes grouped by carcinogenicity among the 4 cytological categories, the HPV genotype found in each sample was coded on an ordinal scale: HPV undetected (0), not classifiable/not carcinogenic (1), possibly carcinogenic (2), carcinogenic (3), and highly carcinogenic, (4). Cytology was also coded on an ordinal scale, NILM (0), ASC-US (1), LSIL (2), and HSIL (3), to determine the correlation between HPV carcinogenicity and cytological grade. For CpG-methylation levels (%), a percentile definition was used. The 95th percentile value for each CpG derived from NILM cytology (HPV-negative) was used as the cut-off for normal methylation (coded as 0); >95th percentile was deemed hypermethylated (coded as 1).

Preceding logistic regression modeling, missing data from the explanatory variables were handled by chained multiple imputation, which fills in missing values of multiple variables iteratively by using chained specifications of prediction equations based on the distribution of each variable (StataCorp, 2019). Multivariable logistic regression (Long and Freese, 2014) was performed to determine the association between the methylation level of each CpG locus of 3 genes (*ADCY8*, *CDH8*, and *ZNF582*) and a binarized cytological outcome of interest. Outcome Model 1 aimed to distinguish normal (NILM) from abnormal (ASC-US/LSIL/HSIL) cytology, Model 2 distinguished NILM/ASC-US from LSIL/HSIL cytology, and Model 3 distinguished HSIL cytology from all others (NILM/ASC-US/LSIL). The covariates (CpG site(s) selected from each gene) that had the highest association with the response variable (p-value < 0.05) were entered in a 2nd multivariable logistic regression jointly with HPV carcinogenicity to select the explanatory variables most predictive of the cytological outcome. The 2nd model equation is as follows:

$$\text{Logistic model}: \text{Probability of outcome} = P(Y = 1)$$
$$= 1/1 + e^{\wedge}[-(b_0 + b_1X_1 + \cdots + b_4X_4)]$$

Multiple explanatory variables: $X_1, \ldots, X_4$.
X_1 = HPV carcinogenicity (coded as ordinal data as described above).
X_2 = *ADCY8* CpG-site *i* methylation (0, 1).
X_3 = *CDH8* CpG-site *i* methylation (0, 1).
X_4 = *ZNF582* CpG-site *i* methylation (0, 1).
Binarized Model 1 Outcome (Y) coding: NILM (0), ASC-US/LSIL/HSIL (1).
Binarized Model 2 Outcome (Y) coding: NILM/ASC-US (0), LSIL/HSIL (1).
Binarized Model 3 Outcome (Y) coding: NILM/ASC-US/LSIL (0), HSIL (1).

For the final regression models, post estimation receiver operating characteristic (ROC) curves were constructed and predictions at specified values were computed. After estimating

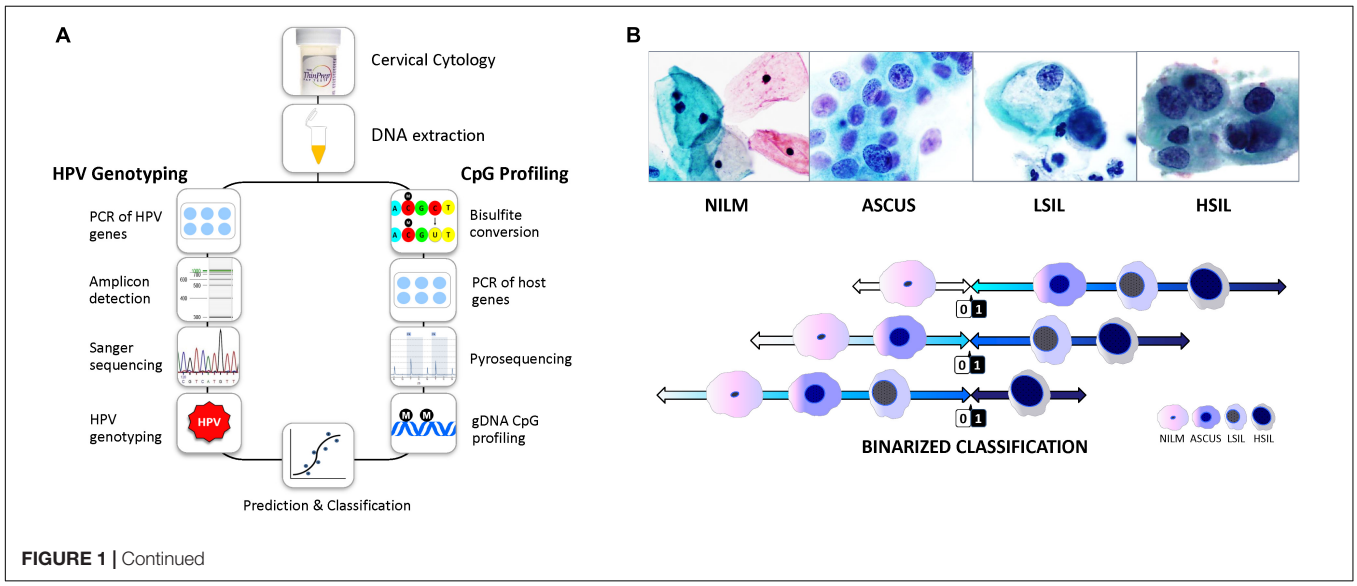


FIGURE 1 | Continued

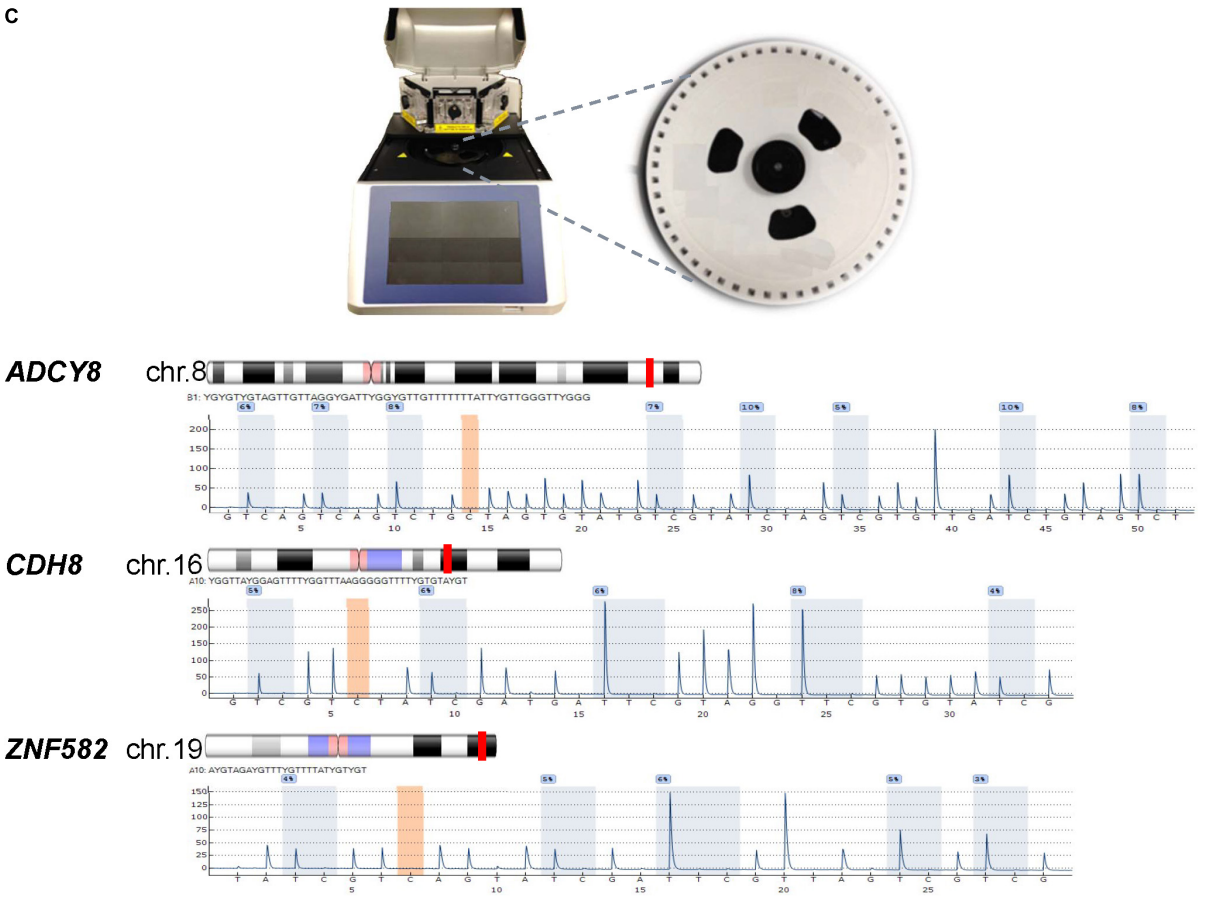


FIGURE 1 | Protocol schema and representative images of four cervical cytological grades used in the study. **(A)** Sample collection, DNA extraction, HPV genotyping by Sanger sequencing and genomic CpG profiling of loci-specific promoters by PSQ. Sequencing results are used for statistical modeling, prediction and classification. **(B)** Four categories of liquid-based cervical cytology: NILM, ASC-US, LSIL, and HSIL with cytomorphologic features of disease progression, i.e., increased nuclear enlargement, nuclear membrane irregularity, nuclear/cytoplasmic ratio, and chromatin coarseness (ThinPrep Pap smear, 50x magnification). Bottom, binarized classification of 4 cytological grades used as outcomes ("0" and "1") for logistic regression. Three distinct, sequential logit models were used for outcome prediction by molecular signatures. **(C)** PyroMark Q48 PSQ instrument and 48-well sample disk (expanded) used for DNA methylation analysis. Bottom, the PSQ CpG assays for three host genes: *ADCY8, CDH8, and ZNF582* are shown according to chromosomal locations (red line). The representative pyrograms with assay specific CpG sites (blue-gray columns) are shown with sequence specific, light-intensity peaks along the x- and y-axis, respectively. Chromosome ideograms adapted from NCBI Map Viewer (www.ncbi.nlm.nih.gov/genome/guide/human)]. ASC-US, atypical squamous cells of undetermined significance; chr, chromosome; gDNA, genomic DNA; HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; NILM, negative for intraepithelial lesion or malignancy; PCR, polymerase chain reaction; PSQ, pyrosequencing. Photo credit (cytology): Bradie Bishop, MD.

the classification threshold or cut-off for each model by using Youden's index (maximum sum of sensitivity and specificity), diagnostic performance characteristics were determined (Youden, 1950). Of note, a cut-off may be adjusted for greater or lesser sensitivity (true positives) while trading off 1-specificity (false positives) based on relative importance of the parameters and purpose of a clinical test. The discriminatory performance between multivariable and univariable (HPV carcinogenicity) models was compared using respective areas under the ROC curve. Pairwise comparisons of predicted probabilities between models were performed with the chi-square test.

Statistical Analysis

This study was designed to have a 90% power to detect a 10% difference in DNA methylation (%) between successive categories of cytology. From the literature, locus-specific promoter methylation levels (%) for NILM, LSIL/HSIL and cervical cancer have ranged from 0–5%, 15–30%, and 30–60%, respectively (Lai et al., 2008; Wentzensen et al., 2009; Siegel et al., 2015). To detect a 10% difference in methylation levels using a one-sided test set at $\alpha = 0.05$ and $\beta = 0.10$ with an allocation ratio of 1, a total accrual target of $N = 306$ and $n = 153$ per group was required. The quota sampling strategy assured adequate representation from each cytological grade. Additional samples were collected to compensate for potential sample inadequacy and laboratory errors.

Data were summarized using means (95% CI), medians (IQR), and proportions. For hypothesis testing, Wilcoxon rank sum and Kruskal-Wallis tests were used for non-parametric, numerical, or ordinal data. Categorical data were compared using

the chi-square test. Correlation between ordinal variables was determined by Spearman's rho. p-values < 0.05 were considered statistically significant. Statistical analyses were performed using STATA/IC 16.0 (StataCorp LP).

RESULTS

HPV Type-Specific Carcinogenicity Correlates With Cytological Grade

A total of 883 residual cytology samples were collected between September 2015 and March 2017. Clinical and cytological characteristics are summarized in **Table 1**. The corresponding subjects were composed predominantly of Caucasians (38%) with a median age of 30 years (IQR, 25–37). The cytological specimens were stratified proportionately among the 4 grades except for HSIL with fewer samples (chi-square $p < 0.05$): NILM 237/883 (27%); ASC-US 235/883 (26%); LSIL 229/883 (26%), and HSIL 181/883 (21%). Sample #503 was classified as HSIL/SCC and the 66-year old patient was subsequently diagnosed of Stage IIA invasive SCC. This sample thus served as the positive control for this study. The median concentrations of extracted DNA among the 4 cytological categories were adequate and statistically equivalent (range, 91.6–119.2 ng/uL) (Kruskal-Wallis test, $p = 0.519$) (**Table 1**).

HPV prevalence determined by PCR and gel electrophoresis increased significantly with worsening grade: NILM (53%), ASC-US (58%), LSIL (97%), and HSIL (86%) (**Figure 2A**). Similarly, the proportion of carcinogenic HPV genotypes in PCR-positive sequenced samples ($n = 640$) increased coincidently with cytological grade: NILM (27%), ASC-US (37%), LSIL (41%), and HSIL (82%) (**Figure 2B**). Conversely, a significant downtrend was found for HPV genotypes in possibly, not carcinogenic/unclassified HPV-types, and HPV-types that were unidentifiable by BLAST ($p < 0.05$, chi-square trend test).

The HPV genotype distribution of 577 sequenced cytology samples spanned the continuum of IARC-defined carcinogenic potentials (**Figure 2C**). The remaining 63 HPV PCR-positive samples could not be genotyped by BLAST due to uninterpretable (poor or noisy) sequencing results. Species richness diminished remarkably with progression of cytopathology from NILM to HSIL (38 to 23 genotypes, respectively). HPV16 surged and dominated the HSIL viral community in 68/152 (45%) samples. Meanwhile, the 12 other carcinogenic and a few possibly carcinogenic and not carcinogenic/unclassified genotypes dwindled but persisted in HSIL samples.

A representative Neighbor-Joining tree constructed from reference sequences of 57 HPV genotypes (one for each genotype identified in the 577 sequenced samples) is presented in **Figure 3**. The concatenated sequences grouped likewise to the conventional L1-based and joined E7/E1/E2/L2/L1-based phylogenetic trees) (International Agency for Research on Cancer [IARC], 2012; Schiffman et al., 2009). The tree revealed an inverse relationship between genetic distance from HPV-16 (highest carcinogenic potential) and carcinogenic risk which corresponded to the prevalent HPV genotypes found among the four grades of cytology as shown in **Figure 2C**.

TABLE 1 | Clinical and cytological characteristics of the study population.

Characteristics	N	(%)[h]
Clinical		
Age[a]		
Median (IQR)	30	(25–37)
Range (year)	20–69	
Race/Ethnicity[a]		
Asian [NILM, ASC-US, LSIL, HSIL]	32 [8, 8, 10, 6]	(3.6)
Black [NILM, ASC-US, LSIL, HSIL]	106 [28, 29, 35, 14]	(12.0)
White [NILM, ASC-US, LSIL, HSIL]	333 [84, 96, 82, 71]	(37.7)
Unknown [NILM, ASC-US, LSIL, HSIL]	407 [115, 102, 99, 91]	(46.1)
Missing [NILM, ASC-US, LSIL, HSIL]	5 [2, 0, 3, 0]	(0.6)
Cytological[b]		
Total LBC samples collected	883	(100)
LBC samples missing clinical data[c]	5	(0.6)
NILM	2	(0.2)
ASCUS	0	(0)
LSIL	3	(0.4)
HSIL	0	(0)
LBC samples included	883	(100)
NILM	237	(27)
ASCUS	235	(26)
LSIL	229	(26)
HSIL	181	(21)
HSIL/SCC	1	(0.1)
Source[d]		
Cervical	875	(99.1)
Vaginal	3	(0.3)
Unspecified	5	(0.6)
Diagnostic category[d]		
Normal	237	(27)
Abnormal	646	(73)
Cellular DNA concentration[d,e,f]		
Total LBC samples [Median (ng/uL) (IQR)]	95.6	(54.3–168.4)
NILM [Median (ng/uL) (IQR)]	91.6	(54.2–147.3)
ASCUS [Median (ng/uL) (IQR)]	81.1	(34.9–147.3)
LSIL [Median (ng/uL) (IQR)]	119.2	(67.5–202.0)
HSIL [Median (ng/uL) (IQR)]	104.0	(54.8–178.2)

ASC-US, atypical squamous cells of undetermined significance; IQR, interquartile range; LBC, Liquid-based Cytology; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion; NILM, negative for intraepithelial lesion and malignancy; SCC, squamous cell carcinoma. [a]Age and Race/Ethnicity of subjects (N = 883) are based on the demographic data of included samples. The number of samples by race/ethnicity and cytological diagnosis are placed in brackets. The distribution of the 5 categories of race/ethnicity (inclusive of Unknown and Missing data) across the 4 cytological grades were not different (χ^2, $p = 0.354$). [b]Cytopathology results are ascribed to the specimens collected on day of study enrollment. [c]Exclusion criteria included: low cell pellet volume (<200 μL). [d]Data based on all included samples (N = 883). Unspecified source (cervix or vagina). [e]Concentration of total cellular DNA per sample after semi-automated DNA extraction. [f]Comparison of DNA concentrations between NILM, ASC-US, LSIL, and HSIL sample groups showed slightly lower concentration for NILM vs. LSIL, ASC-US vs. LSIL, and ASC-US vs. HSIL ($p < 0.05$) by the Wilcoxon rank-sum test. [h]Values are N (%) unless otherwise denoted.

Promoter Hypermethylation of ADCY8, CDH8, and ZNF582 Correlates With Cytological Grade

Methylation levels for all CpG sites increased coincidently with worsening cytological grade for *ADCY8, CDH8,* and

ZNF582 except for *ADCY8* at CpG sites 1-5 (Spearman rank, $p < 0.05$). Pairwise comparisons of methylation for each CpG site between cytological grades (NILM vs. ASC-US, ASC-US vs. LSIL, and LSIL vs. HSIL) revealed significantly higher levels at multiple sites for the worse grade denoted by a star in **Figure 4** (*$p < 0.05$ by the Wilcoxon rank-sum test). The upper limit of normal for CpG-methylation levels summarized as the median of the 95th percentile of each CpG site per assay were: *ADCY8* (10.11%), *CDH8* (7.61%), and *ZNF582* (5.22%). Positive control sample #503 bore methylation levels 2-fold and 3-fold that of NILM cytology for *ADCY8* and *CDH8/ZNF582* assays, respectively (**Supplementary Figure 3**). CpG assay results

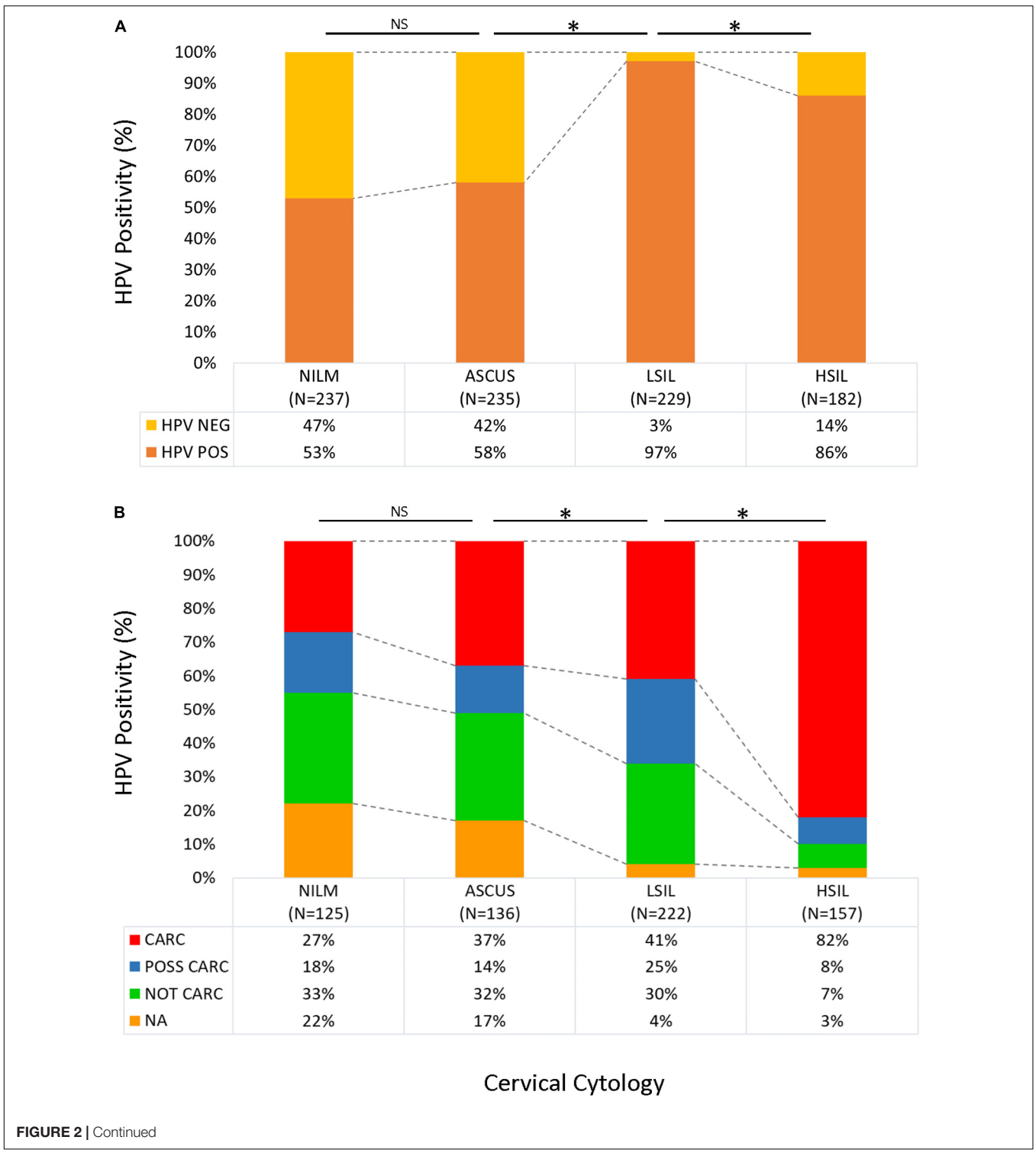


FIGURE 2 | Continued

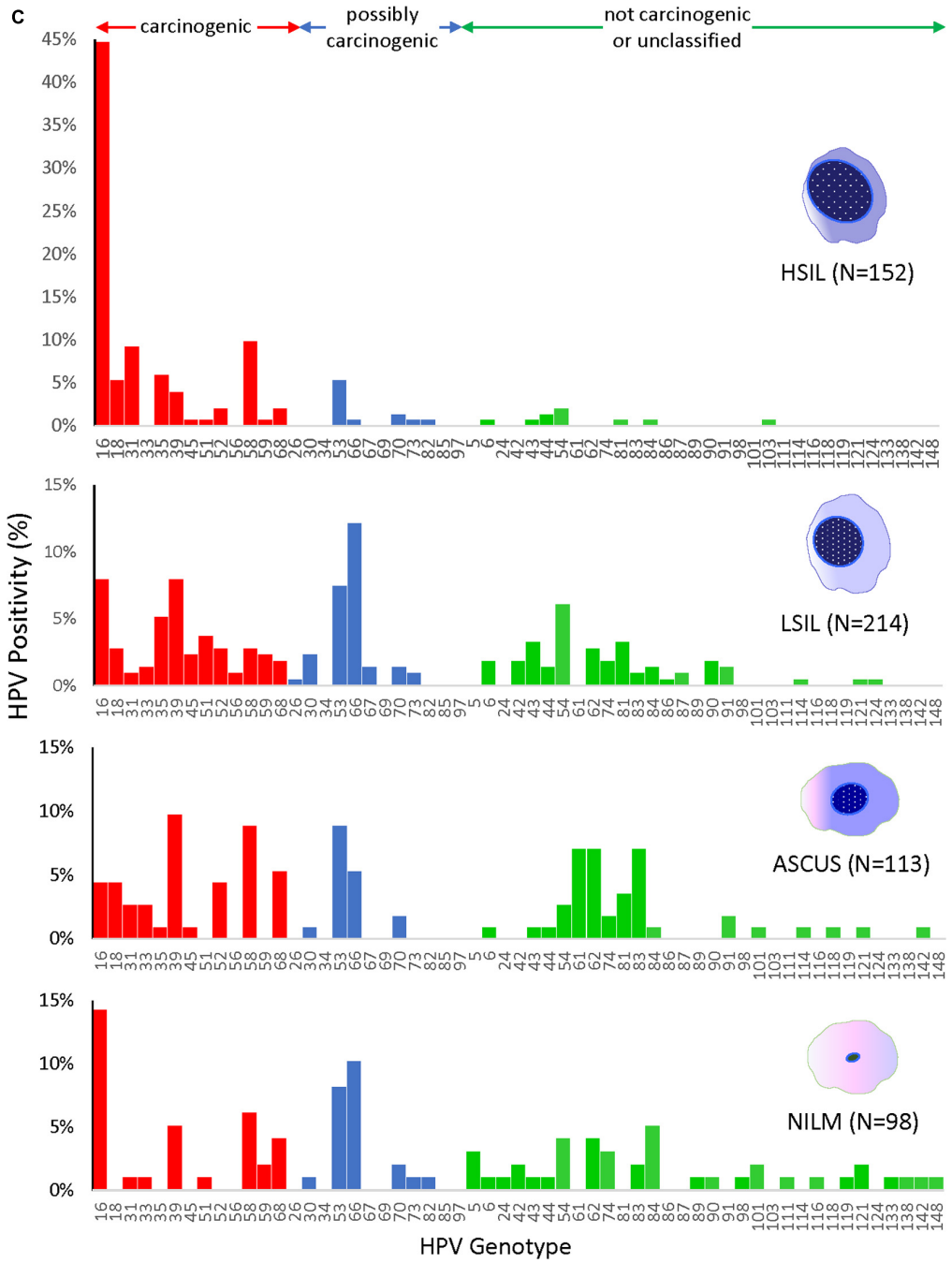


FIGURE 2 | HPV prevalence and genotype distribution found in 4 cytological grades. **(A)** HPV DNA positivity rate for 883 samples as determined by PCR amplification and gel electrophoresis. The positive rates for NILM and ASCUS were over 50%, whereas the rates were significantly higher for LSIL and HSIL at ~80–90% (top) (*$p < 0.05$, chi-square test). **(B)** Distribution of HPV-positive rates stratified by type-specific carcinogenic potential for PCR-positive/sequenced samples ($n = 640$). Progression of cytological grade from NILM to HSIL correlated with a significant uptrend in carcinogenic HPV-types and a downtrend in possibly and not carcinogenic/unclassified HPV-types (*$p < 0.05$, chi-square trend test). Samples with poor or noisy sequence quality unidentifiable by BLAST also decreased with worsening cytological grade (*$p < 0.05$, chi-square trend test). **(C)** HPV genotype distribution of 577 cytology samples as determined by PCR/Sanger sequencing according to cytological diagnoses. The remaining 63 HPV-positive samples could not be genotyped due to poor sequence quality and/or overlapping sequences of mixed infections. The proportion of carcinogenic HPV genotypes (red bars) increased coincidently with cytological grade (*$p < 0.05$, chi-square trend test). In contrast, the possibly and not carcinogenic/unclassified HPV-types (blue and green bars, respectively) significantly diminished (*$p < 0.05$, chi-square trend test). Simultaneously, species richness diminished from NILM to HSIL (38 to 23 genotypes, respectively) while HPV-16 surged in 68/152 (45%) HSIL samples. ASC-US, atypical squamous cells of undetermined significance; CARC, carcinogenic HPV; HSIL, high- grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; NA, not available/identifiable by BLAST; NILM, negative for intraepithelial lesion or malignancy; NOT CARC, not carcinogenic; NS, not significant; POSS CARC, possibly carcinogenic Stars, $p < 0.01$.

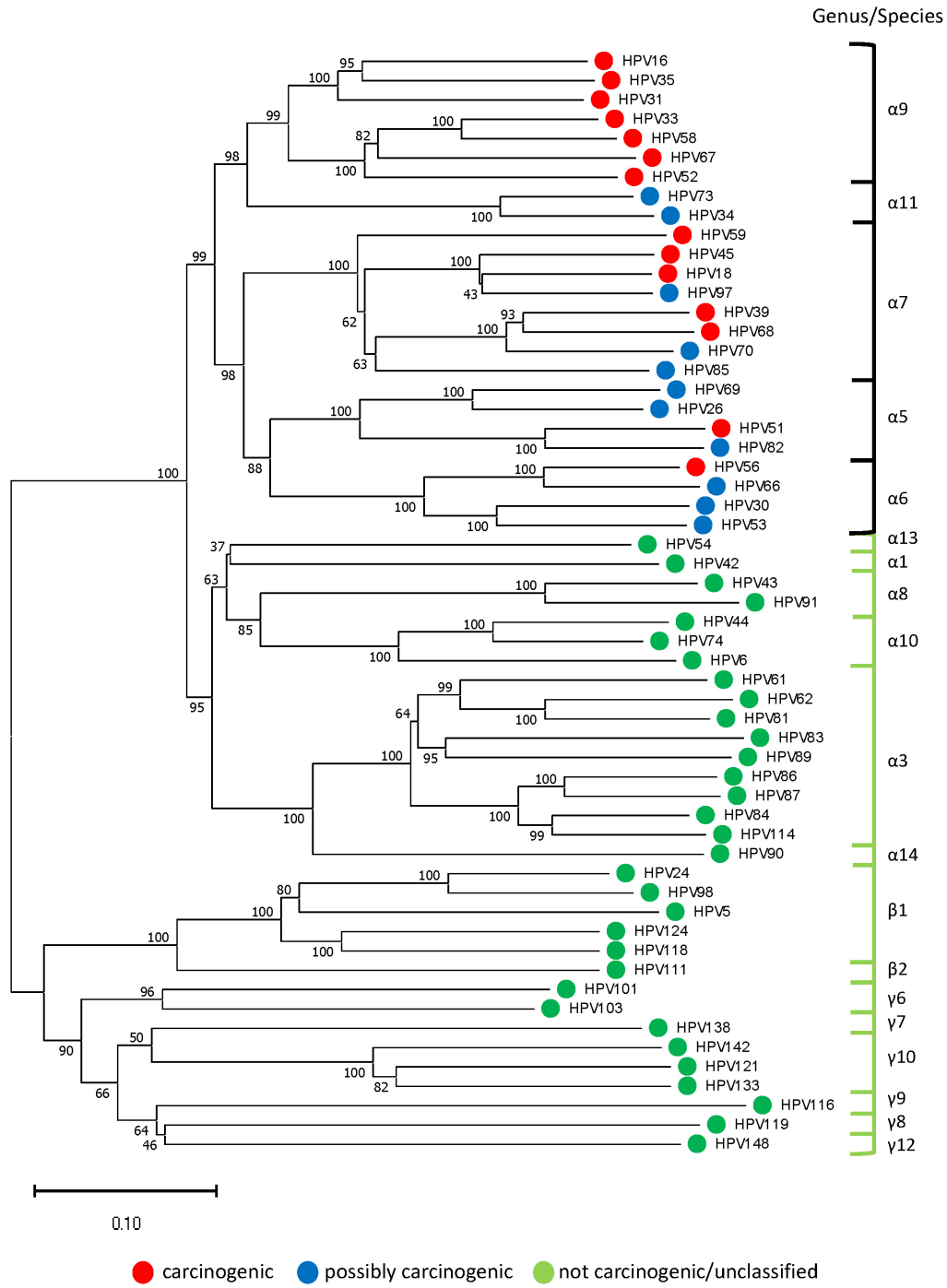


FIGURE 3 | Representative phylogenetic tree of HPV genotypes identified in the clinical samples. Neighbor-Joining tree of 57 HPV genotypes (one from each genotype) revealed two distinct clades in the *alpha* genera: "high-risk" containing carcinogenic and possibly carcinogenic types [black bracket] and "low-risk" containing probably not carcinogenic or not classifiable types [green bracket]. The *beta and gamma* genera formed another clade composed of commensal and unclassified genotypes. With HPV-16 at the pinnacle of HPV carcinogenic potential, genetic divergence from this point correlated with decreased carcinogenic risk (phenotype) and grade of cytopathology. The evolutionary history was inferred using the Neighbor-Joining method after concatenating 57 aligned, E6, E7, and L1 reference nucleotide sequences from Papillomavirus Episteme by MUSCLE (Edgar, 2004). The optimal tree with the sum of branch length = 9.99536245 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004) and are in the units of the number of base substitutions per site. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 2850 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018).

analyzed on the PyroMark Q48 were also compared to Q96 data from our previous study (Shen-Gunther et al., 2016). This revealed a slightly higher CpG-methylation level for each site on the Q48 indicative of improved analytical sensitivity (**Supplementary Figure 4**).

ADCY8, CDH8, and ZNF582 CpG-Markers Contribute to HPV as a Predictor of Cytological Grade

ROC curve analysis after multivariable logistic regression for three logit models are presented in **Figure 5**. The best predictors for differentiating NILM from ASC-US/LSIL/HSIL was HPV carcinogenicity, *ZNF582*_1st CpG site, and *ADCY8*_5th CpG site (**Supplementary Table 1**). The best multivariable predictor for differentiating between NILM/ASC-US and LSIL/HSIL cytology was the combination of HPV carcinogenicity, *ZNF582*_1st CpG site, *CDH8*_4th CpG site, and *ADCY8*_5th CpG site (**Supplementary Table 2**). Additionally, the combination of HPV carcinogenicity, *ZNF582*_1st CpG site and *CDH8*_4th CpG site, and *ADCY8*_6th CpG site is the best multivariable predictor for differentiating between NILM/ASC-US/LSIL and HSIL cytology (**Supplementary Table 3**). All three multivariable models were better predictors of the specified outcome than HPV carcinogenicity alone (delta AUC*, $p < 0.05$, chi-square test). Overall, the number of missing observations for HPV genotype 63/883 (7.1%) and site-specific CpG-methylation ranging from 1 to 9/883 ($\leq$ 1%) were low prior to imputation and predictive modeling. The variables with counts of missing observations are listed in **Supplementary Table 4**.

HPV With Host CpG-Markers Outperforms HPV as a Classifier of Cervical Cytology

Predicted probabilities plots of binarized cytological outcomes using HPV carcinogenicity as a singular or integrated predictor variable are shown in **Figure 6**. HPV carcinogenicity as a one-dimensional predictor of 3 sequentially binarized cytological outcomes (NILM vs. ASC-US/LSIL/HSIL, NILM/ASC-US vs. LSIL/HSIL, and NILM/ASC-US/LSIL vs. HSIL) are shown with respective cut-off values of $\geq$0.680, 0.5222, and 0.3321 as determined by Youden's index (**Figure 6A**; Youden, 1950). HPV-16 was distinct from the other carcinogenic HPVs in predicting the cytological outcome of interest.

HPV carcinogenicity and host loci-specific methylation as predictors of cytological outcome are shown in **Figure 6B**. The 3 subgraphs illustrate the escalating probability for the cytological outcome of interest coincident with HPV carcinogenicity and increased number of methylated genes. For example, the probability for HSIL escalated with HPV carcinogenicity and increased counts of methylated genes reaching the pinnacle of 96% at HPV-16 jointly with 3 methylated genes. Furthermore, the contribution of each methylated gene to the probability of the outcome of interest are different singularly or in combination with others as shown by non-overlapping lines (**Figure 6B**). The cut-off values for predicting positive binarized cytological outcomes (NILM vs. ASCUS/LSIL/HSIL, NILM/ASCUS vs. LSIL/HSIL, and NILM/ASCUS/LSIL vs. HSIL) were $\geq$0.6503, 0.4533, and 0.2645 as determined by Youden's index (Youden, 1950).

The diagnostic performance characteristics of the three logit models were presented in **Supplementary Tables 5–7**. Specifically, models 1 and 2 inclusive of methylation markers showed greater sensitivity (83% and 79%, respectively) and a lower negative likelihood ratio compared to HPV alone. This implies that HPV with methylation markers performed better at predicting absence of disease (more true negative results) and in differentiating normal (NILM) from abnormal (>NILM) in model 1 as well as NILM/ASC-US from LSIL/HSIL in model 2. In contradistinction, model 3 performed better at predicting presence of disease (more true positive results) with greater specificity (88%) and had a higher positive likelihood ratio of 5.55. In other words, a positive result was ∼5 times more likely to occur in a patient with HSIL than one with < HSIL.

DISCUSSION

This study aimed to validate the predictive model of a multiparametric biomarker panel based on HPV genotype and host epigenetic modifications for cervical cytopathology. The expanded results of the current study confirmed our previous findings, which determined that HPV carcinogenicity is positively correlated with aberrant DNA methylation in the promoters of *ADCY8*, *CDH8*, and *ZNF582* in addition to the four cytological grades (Shen-Gunther et al., 2016). In this study, the addition of ASC-US samples revealed a distinct and intermediate HPV genotype prevalence pattern between NILM and LSIL. Similarly, the promoter methylation levels of ASC-US were found between that of NILM and LSIL. Overall, the multivariate biomarker panel improved the prediction and classification of cytological grade over the univariate HPV carcinogenicity.

The HPV genotype patterns among the cytological categories revealed a loss of species diversity and gain of dominance by carcinogenic types as cytology progressed from NILM to HSIL. Our PCR results showed a doubling in HPV positivity rate from NILM/ASC-US to LSIL/HSIL samples (>80–90%). PCR-sequencing revealed a steep rise in the proportion of carcinogenic genotypes as cytology progressed to HSIL with the 3 topmost being HPV-16 (45%), HPV-58 (10%), and HPV-31 (9%). In contrast, the possibly carcinogenic and non-carcinogenic HPV genotypes declined precipitously in prevalence and diversity indicative of extinction of the "weaker competitors" within the dynamic virus-host ecosystem. This phenomenon, demonstrated previously by our deep sequencing study, may be explained by the principles of competitive exclusion followed by clonal expansion of HPV-infected transformed cells (Ueda et al., 2003, 2008; Shen-Gunther et al., 2017). The competitive advantage of carcinogenic, particularly HPV-16, versus non-carcinogenic HPV is derived from the significantly higher binding affinity of viral E6 and E7 oncoproteins to host cellular proteins p53 and pRB, respectively, leading to subsequent degradation, genomic instability, and carcinogenesis (Chemes et al., 2010; Martinez-Zapien et al., 2016;

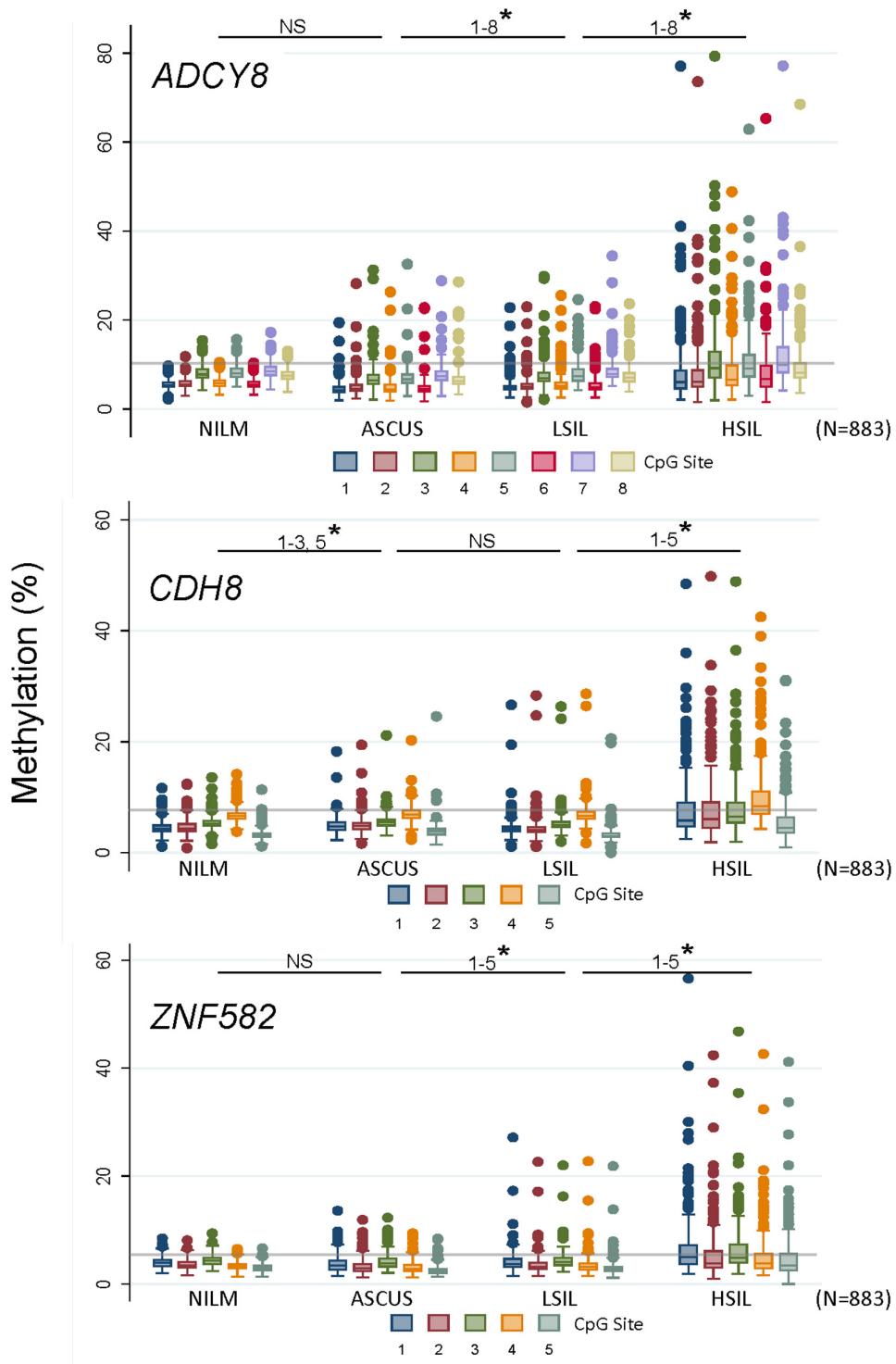


FIGURE 4 | Loci-specific promoter methylation differences and trends among cervical cytological grades. Methylation (%) of total genomic DNA in 4 grades of cervical cytology i.e., NILM (n = 237), ASC-US (n = 235), LSIL (n = 229), and HSIL (n = 182) was compared by CpG sites among 3 genes (*ADCY8*, *CDH8*, and *ZNF582*). Pairwise comparisons of methylation for each CpG site between cytological grades (NILM vs. ASC-US, ASC-US vs. LSIL, and LSIL vs. HSIL) revealed significantly higher levels at multiple sites as noted by a star (* $p < 0.05$ by the Wilcoxon rank-sum test). Methylation levels for all CpG sites increased coincidently with cytological grade for *ADCY8*, *CDH8*, and *ZNF582* by Spearman's r_s ($p < 0.05$, with Bonferroni adjustment) except for *ADCY8* CpG sites 1–5. The methylation reference line (gray) for each assay denotes the median of the 95th percentile values for each CpG site within an assay derived from NILM (HPV-negative) samples, i.e., *ADCY8* (10.11%), *CDH8* (7.61%), and *ZNF582* (5.22%). ASC-US, atypical squamous cells of undetermined significance; HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; NILM, negative for intraepithelial lesion or malignancy; NS, not statistically significant.

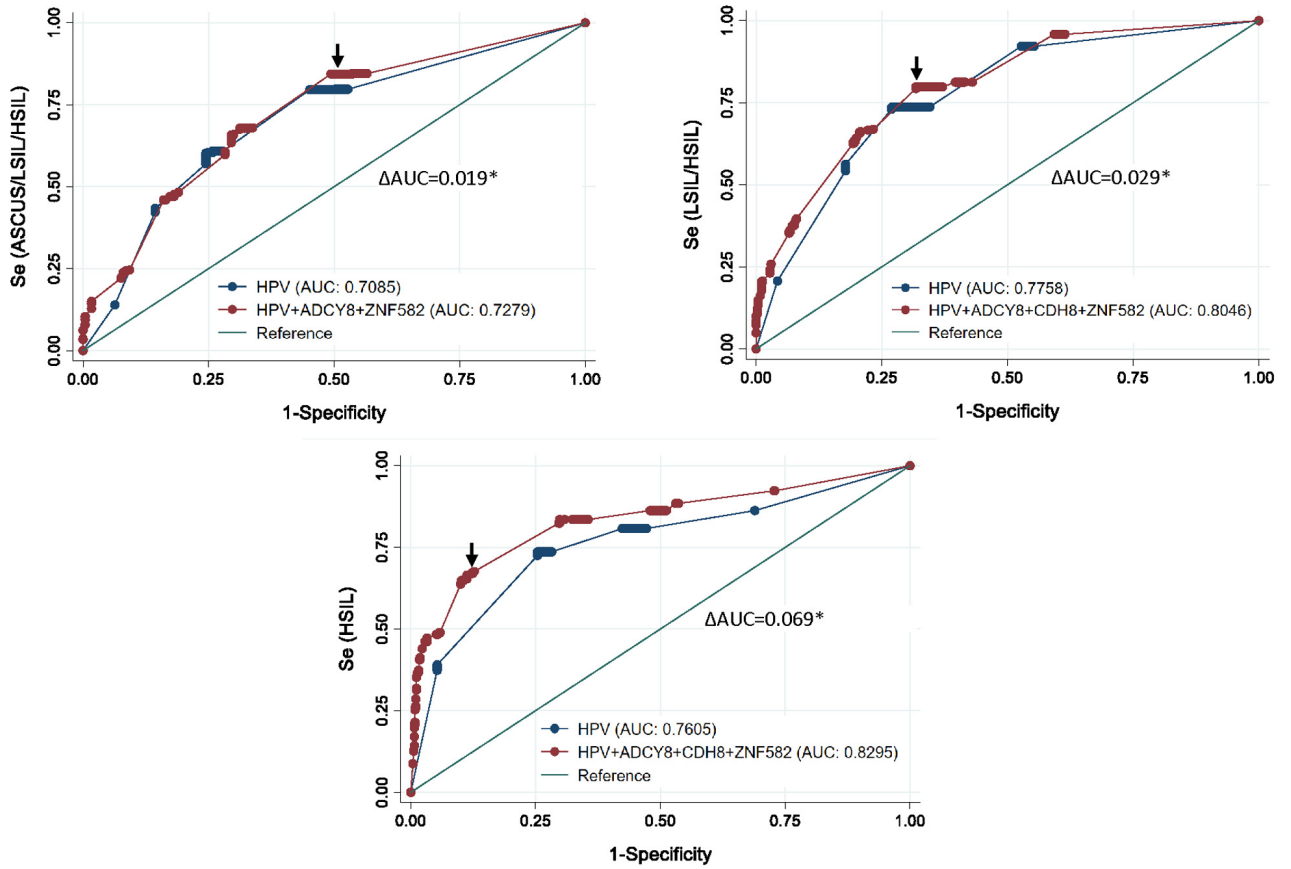


FIGURE 5 | Receiver operating characteristic (ROC) curve analysis after multivariable logistic regression for three logit models. Top left, the ROC curve revealed the best predictors to differentiate between NILM and ASCUS/LSIL/HSIL as HPV carcinogenicity, *ZNF582*_1st CpG site, and *ADCY8*_5th CpG site. Top right, for differentiating between NILM/ASCUS and LSIL/HSIL cytology, the best multivariate predictor was the combination of HPV carcinogenicity, *ZNF582*_1st CpG site, *CDH8*_4th CpG site, and *ADCY8*_5th CpG site. Bottom, for differentiating between NILM/ASCUS/LSIL and HSIL cytology, the best multivariate predictor was the combination of HPV carcinogenicity, *ZNF582*_1st CpG site, *CDH8*_4th CpG site, and *ADCY8*_6th CpG site. All three multivariable models were better predictors of the specified outcome than HPV carcinogenicity alone (delta AUC*, $p < 0.05$, chi-square test). AUC, Area under the receiver operator curve (AUC); se, sensitivity; cut-off points (arrows).

Tomita et al., 2020; ðukić et al., 2020). Furthermore, the distinctive cellular binding motifs of E6 and E7 oncoproteins, which correlate with tissue and host specificity, is believed to be the result of virus-host adaptive evolution over millions of years (Chemes et al., 2012; Chen et al., 2018; Suarez and Trave, 2018). Evolutionary analysis of 57 representative genotypes confirmed the inverse relationship between genetic distance from HPV-16 and carcinogenic potential. As a corollary, the type-specific carcinogenic risk was reflected in the severity of cytopathology.

Cellular epigenetic analyses confirmed our previous findings of a positive correlation between promoter methylation of *ADCY8*, *CDH8*, and *ZNF582* and cytological grade (Shen-Gunther et al., 2016). The curvilinear upward trend in quantitated methylation levels for all 3 genes was validated on the high-resolution PyroMark Q48 instrument. Our results are consistent with E6 and E7 oncoprotein induced promotion of *de novo* and/or maintenance DNA methyltransferase (DNMT1, DNMT3A, and DNMT3B) (Burgers et al., 2007; Au Yeung et al., 2010; Durzynska et al., 2017). The loci-specific methylation profiles from the Q48 mirrored that of Q96 from our previous study (Shen-Gunther et al., 2016). However, CpG-site specific median methylation levels were consistently slightly higher for Q48 which ranged from 1 to 8%. We ascribe this finding to the advanced chemistry and improved algorithms employed in the Q48 resulting in reduced background and augmented sensitivity in sequencing reactions (Qiagen, 2020). High-resolution PSQ offered a technological advantage in this study by exposing subtle differences in methylation levels between gene specific CpG sites and identifying the CpG site that contributed most to our predictive models. Along this line, Lioznova and coauthors dubbed unique single CpG methylation sites as "CpG traffic lights" which were found to correlate more often with gene expression and repression than an averaged promoter methylation statistic (Medvedeva et al., 2014; Lioznova et al., 2019).

Model refitting was performed due to sample size expansion, ASC-US sample inclusion, and PSQ technological advancement in this cohort. The combination of HPV carcinogenicity

and methylation markers remained significant predictors of cytological outcome after multivariable logistic regression. The addition of methylation status improved the sensitivity and/or specificity for the binarized outcome of interest for all 3 models. The predicted probability for the outcome escalated in a stepwise fashion as HPV carcinogenicity reached its pinnacle, i.e., HPV-16. This finding is consistent with a longitudinal study where HPV-16, in comparison to other carcinogenic HPV types, was found to convey a uniquely elevated risk for severe cervical intraepithelial neoplasia (CIN3 +) (Demarco et al., 2020). Furthermore, as the number of methylated genes increased, the probability for the outcome increased. Whereas slight differences were noted between specific genes. Together, infection with the most carcinogenic HPV and maximal loci methylation predicted the greatest probability for high-grade cytology.

Discriminatory performance of the 3 multi-parametric models demonstrated greater accuracy than the univariate HPV carcinogenicity. Specifically, models 1 and 2 inclusive of methylation markers showed greater sensitivity and a lower negative likelihood ratio. This implies that HPV with methylation markers performed better at predicting absence of disease (more true negative results). Conversely, model 3 performed better at predicting presence of disease (more true positive results) with greater specificity and a higher positive likelihood ratio of 5.55. Likelihood ratios, akin to signal-to-noise ratios, are robust measures of diagnostic accuracy that are independent of disease prevalence (Deeks and Altman, 2004; Lang and Secic, 2006). This statistic allows for generalizability and comparison beyond the scope of this study. Finally, our predictive models and classification schemes designed for sequential use will enable the allocation of an unknown sample to the appropriate cytological category (Long and Freese, 2014).

Published literature on promoter methylation of *ADCY8, CDH8,* and *ZNF582* has expanded recently and hypermethylation of one or more of these loci have been found in cancers of the breast, oropharynx, esophagus, and anus (Tang et al., 2019; Ekanayake Weeramange et al., 2020; Sigin et al., 2020; van der Zee et al., 2020). These reports not only support the validity of these epigenetic modifications as cancer biomarkers but inform a broader application beyond cervical cancer. First, differential methylation of *ADCY8* was identified as one of three most informative biomarkers in luminal B breast cancer from a cohort of Russian women. Methylation status of these markers were predictive of response to neoadjuvant chemotherapy before surgery which may be applicable as a clinical test to guide therapy (Sigin et al., 2020). Second, *CDH8* promoter hypermethylation has been documented in four head and neck cancer studies (Ekanayake Weeramange et al., 2020). One such study used PSQ for validation of epigenetic alteration in 70 oropharyngeal squamous cell carcinomas. HPV-positive, in contrast to HPV-negative tumors, was found to be significantly correlated

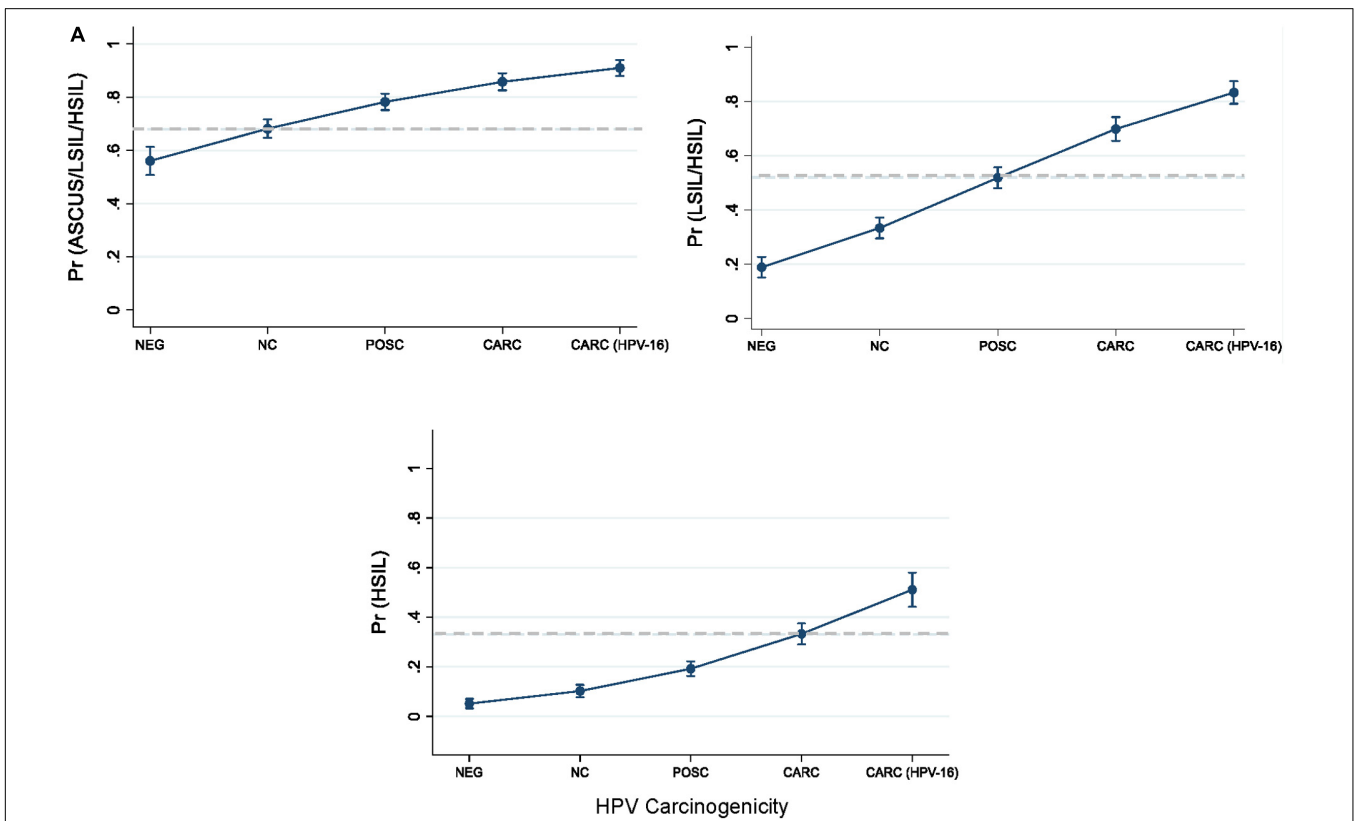


FIGURE 6 | Continued

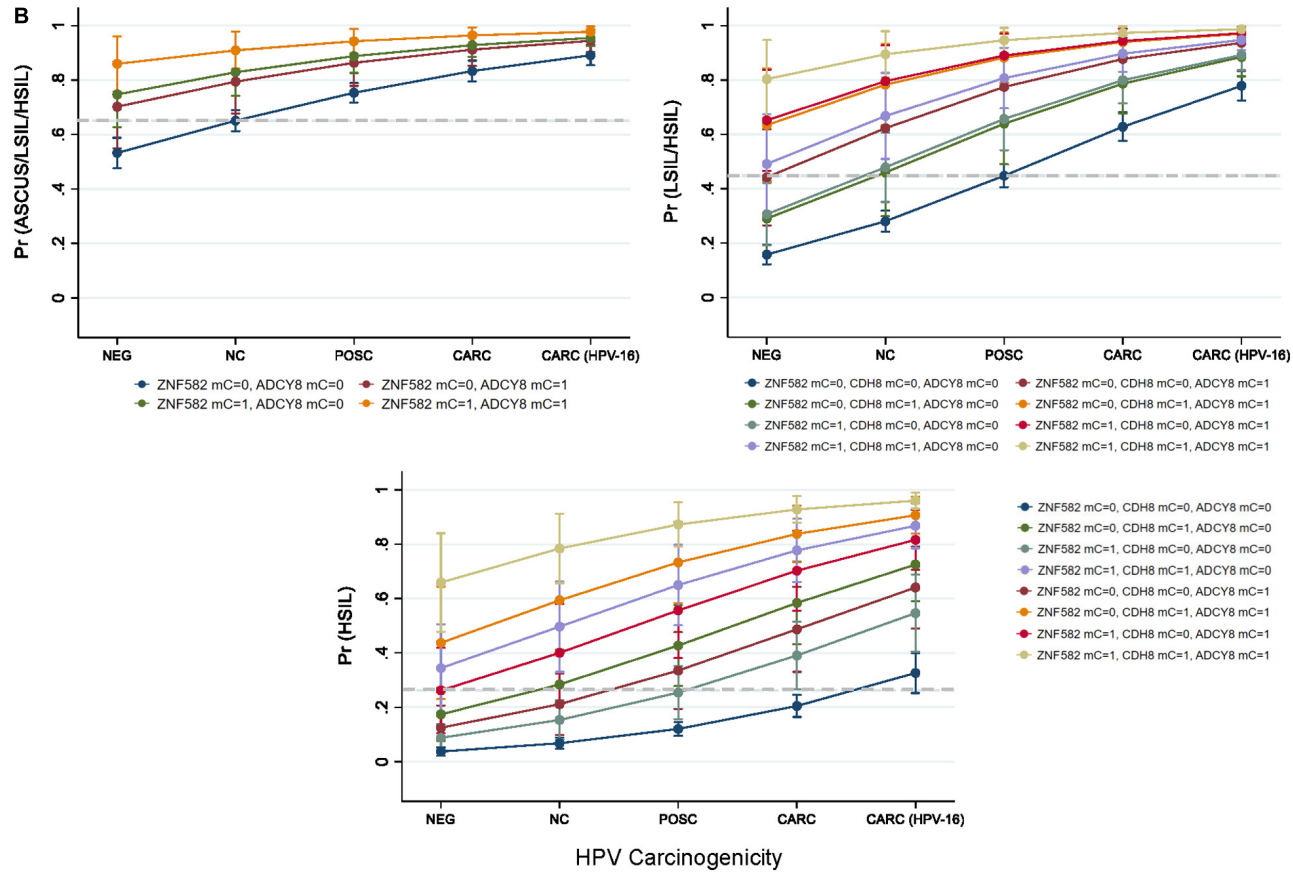


FIGURE 6 | Predicted probabilities plot of binarized cytological outcomes using HPV carcinogenicity as a singular or integrated predictor variable. **(A)** HPV carcinogenicity as a one-dimensional predictor of 3 sequentially binarized cytological outcomes (NILM vs. ASC-US/LSIL/HSIL, NILM/ASCUS vs. LSIL/HSIL, and NILM/ASC-US/LSIL vs. HSIL) is shown with respective cut-off values of ≥0.680, 0.5222, and 0.3321 (dashed lines) as determined by Youden's index. **(B)** HPV carcinogenicity and host loci-specific methylation as predictors of cytological outcome. Top left, comparison of predicted probabilities for abnormal cytology (NILM vs. ASC-US/LSIL/HSIL) by HPV carcinogenicity and binarized *ZNF582* and *ADCY8* methylation status. Top right, comparison of predicted probabilities for NILM/ASC-US vs. LSIL/HSIL permuted by binarized methylation values of *ADCY8, CDH8,* and *ZNF582* at the CpG sites noted in the text. Bottom, comparison of predicted probabilities for <HSIL vs. HSIL permuted by binarized methylation values of *ADCY8, CDH8,* and *ZNF582* at the CpG sites noted in the text. The cut-off values for predicting a positive binarized cytological outcomes (NILM vs. ASC-US/LSIL/HSIL, NILM/ASC-US vs. LSIL/HSIL, and NILM/ASC-US/LSIL vs. HSIL) were ≥0.6503, 0.4533, and 0.2645 (dashed lines) as determined by Youden's index. Definitions: For loci-specific CpG methylation levels (%), the 95th percentile value for each CpG derived from HPV-negative. NILM cytology was used as the cut-off for normal methylation (coded as 0); >95th percentile was deemed hypermethylated (coded as 1). AUC, area under the curve; mC, 5-methylcytosine at CpG sites; mC = 0, unmethylated cytosine; mC = 1, methylated cytosine; Pr, probability; ROC, Receiver operating characteristic; Se, sensitivity.

with hypermethylation and prognosis in this Japanese cohort (Nakagawa et al., 2017).

ZNF582 is the best studied of our 3 methylation markers. *ZNF582* promoter hypermethylation has been confirmed in multiple of studies of cervical precancerous lesions, and invasive adeno- and squamous carcinomas (Huang et al., 2012; Chang et al., 2014; Lai et al., 2014; Lin et al., 2014; Liang et al., 2020). In fact, the analysis of *ZNF582* by methylation-specific quantitative PCR is being commercialized as an *in vitro* diagnostic test (Beltrán-García et al., 2019). Hypermethylated *ZNF582* and *PAX1* genes have also been found in mouth rise samples applicable to the detection of oral dysplasia and cancer (Cheng et al., 2018). In Esophageal Squamous Cell Carcinoma (ESCC), aberrant hypermethylation of *ZNF582* and *PAX1* have been demonstrated using quantitative methylation-specific PCR with levels at 21% versus 0% for tumor and peri-tumor normal tissues, respectively (Huang et al., 2017). Another study of esophageal cancer found significantly higher methylation levels by PSQ in cancerous than adjacent non-cancerous and normal tissues, respectively: 31%, 11%, and 15% (Tang et al., 2019). Finally, in a Netherlands study of 345 anal intraepithelial neoplasia (AIN grades 1-3) and invasive carcinoma samples, *ZNF582* methylation levels escalated with increased disease severity. Among the markers studied, *ZNF582* was the most accurate for detecting AIN grade 3 with immense potential as a clinical biomarker (van der Zee et al., 2020). Taken together, *ADCY8, CDH8,* and *ZNF582* promoter methylation are promising predictive and prognostic biomarkers for multiple tumor types crossing geographic and racial boundaries that undeniably merits further validation.

The strength of this study lies in the inclusion of ASC-US samples and sample size expansion for all cytological categories. This led to increased precision and power of distinction between the four grade-specific HPV communities and methylation levels for model validation. Additionally, high resolution PSQ played a critical role in pinpointing the gene specific CpG that contributed the most to the predictive models, as well as, exposing the subtle differences between CpG sites. PSQ has also been proven as a superior method to methylation-specific PCR for prognostication of survival outcomes (Johannessen et al., 2018). We acknowledge that our study has limitations in that two remaining, uncommon cytological categories with potentially different risk profiles, i.e., ASC-H and AGUS were not included (Nayar et al., 2020). To fill this gap, ASC-H and AGUS samples have been collected for our ongoing large-scale study (>3,000 samples), which is intended to complete our investigation and understanding of molecular evolution within a dynamic virus-host ecosystem.

CONCLUSION

Our expanded findings validated the multivariable prediction model developed for cytological classification. The sequencing-based "Molecular Pap smear" outperformed the singular HPV carcinogenicity in predicting four grades of cervical cytology. Additional host epigenetic markers that evolved with disease progression contributed to the overall classification accuracy.

AUTHOR'S NOTE

This paper has undergone PAO review at Brooke Army Medical Center and was cleared for publication. The view(s) expressed herein are those of the authors and do not reflect the official policy or position of Brooke Army Medical Center, the United States Army Medical Department, the United States Army Office of the Surgeon General, the Department of the Army, the Department of the Air Force, or the Department of Defense or the United States Government.

Presented virtually at the 33rd International Papillomavirus Conference (IPVC 2020) on July 23, 2020.

DATA AVAILABILITY STATEMENT

Primary sequence data will be deposited in the NCBI GenBank upon patent issuance.

ETHICS STATEMENT

This study was approved by the institutional review board of Brooke Army Medical Center, Fort Sam Houston, Texas.

AUTHOR CONTRIBUTIONS

JS-G conceived and designed the study and analyzed and interpreted the data. JS-G, QX, WS, and HA participated in the acquisition of data and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

Laboratory materials for this work were supported in part by the Department of Clinical Investigation Intramural Funding Program at Brooke Army Medical Center, Fort Sam Houston, Texas.

ACKNOWLEDGMENTS

We are grateful to Ms. Roxanne Toscano and Ms. Rosalyn Miller at the Cytopathology Laboratory of the Department of Pathology and Area Laboratories Services (DPALS) and the laboratory staff of the Department of Clinical Investigation (DCI) for their invaluable service in collecting the clinical samples in support of the HPV Research Program at Brooke Army Medical Center. We are also grateful to Bradie Bishop, MD of DPALS for providing the cervical cytology images.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.595902/full#supplementary-material

Supplementary Figure 1 | Representative Hpv genomes and target regions used for Hpv Dna detection by Pcr amplification.

Supplementary Figure 2 | Representative electropherograms and gel images of target amplicons from *Adcy8, Cdh8*, and *Znf582* promoter regions for downstream pyrosequencing.

Supplementary Figure 3 | Pyrograms of *Adcy8*, *Cdh8*, and *Znf582* for sample # 503 with a diagnosis of Hsil/Squamous cell carcinoma on cervical cytology.

Supplementary Figure 4 | Comparison of *Adcy8, Cdh8,* and *Znf582* CpG assay results analyzed on the PyroMark Q48 and Q96 platforms.

Supplementary Table 1 | Logistic regression analysis of Hpv and Hpv + *Adcy8* + *Znf582* for predicting abnormal (Asc-Us/Lsil/Hsil) cytology.

Supplementary Table 2 | Logistic regression analysis of Hpv and Hpv + 3-gene methylation markers for predicting abnormal (Lsil/Hsil) cytology.

Supplementary Table 3 | Logistic regression analysis of Hpv and Hpv + 3-gene methylation markers for predicting for Hsil cytology.

Supplementary Table 4 | Predictor variables used for logistic regression with missing observations.

Supplementary Table 5 | Diagnostic performance of Hpv vs. Hpv + *Adcy8* + *Znf582* for predicting abnormal (Asc-Us/Lsil/Hsil) cytology.

Supplementary Table 6 | Diagnostic performance of Hpv vs. Hpv + 3-gene methylation markers for predicting Lsil/Hsil cytology.

Supplementary Table 7 | Diagnostic performance of Hpv vs. Hpv + 3-gene methylation markers for predicting Hsil cytology.

REFERENCES

Au Yeung, C. L., Tsang, W. P., Tsang, T. Y., Co, N. N., Yau, P. L., and Kwok, T. T. (2010). Hpv-16 E6 upregulation of Dnmt1 through repression of tumor suppressor p53. *Oncol. Rep.* 24, 1599–1604. doi: 10.3892/or_00001023

Beerenwinkel, N., Greenman, C. D., and Lagergren, J. (2016). Computational cancer biology: an evolutionary perspective. *PLoS Comput. Biol.* 12:e1004717. doi: 10.1371/journal.pcbi.1004717

Beltrán-García, J., Osca-Verdegal, R., Mena-Mollá, S., and García-Giménez, J. L. (2019). Epigenetic IVD tests for personalized precision medicine in cancer. *Front. Genet.* 10:621. doi: 10.3389/fgene.2019.00621

Bernard, H. U., Burk, R. D., Chen, Z., van Doorslaer, K., zur Hausen, H., and de Villiers, E. M. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401, 70–79. doi: 10.1016/j.virol.2010.02.002

Bosch, F. X., Broker, T. R., Forman, D., Moscicki, A. B., Gillison, M. L., and Doorbar, J. (2013). 'Comprehensive Control of HPV Infections and Related Diseases' Vaccine Volume 30, Supplement 5, 2012. Comprehensive control of human papillomavirus infections and related diseases. *Vaccine* 31(Suppl. 8), I1–I31. doi: 10.1016/j.vaccine.2013.07.026

Buck, C. B., Day, P. M., and Trus, B. L. (2013). The papillomavirus major capsid protein L1. *Virology* 445, 169–174. doi: 10.1016/j.virol.2013.05.038

Burgers, W. A., Blanchon, L., Pradhan, S., de Launoit, Y., Kouzarides, T., and Fuks, F. (2007). Viral oncoproteins target the DNA methyltransferases. *Oncogene* 26, 1650–1655. doi: 10.1038/sj.onc.1209950

Canfell, K., Kim, J. J., Brisson, M., Keane, A., Simms, K. T., Caruana, M., et al. (2020). Mortality impact of achieving WHO cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries. *Lancet* 395, 591–603. doi: 10.1016/S0140-6736(20)30157-4

Carmichael, D. E., and Cameron, C. (1973). "The Pap smear," in *The Pap Smear: Life of George*, ed. N. Papanicolaou (Springfield: Charles C. Thomas), 68–83.

Chang, C. C., Huang, R. L., Wang, H. C., Liao, Y. P., Yu, M. H., and Lai, H. C. (2014). High methylation rate of LMX1A, NKX6-1, PAX1, PTPRR, SOX1, and ZNF582 genes in cervical adenocarcinoma. *Int. J. Gynecol. Cancer* 24, 201–209. doi: 10.1097/IGC.0000000000000054

Chemes, L. B., Glavina, J., Faivovich, J., de Prat-Gay, G., and SÃnchez, I. E. (2012). Evolution of linear motifs within the papillomavirus E7 oncoprotein. *J. Mol. Biol.* 422, 336–346. doi: 10.1016/j.jmb.2012.05.036

Chemes, L. B., Sánchez, I. E., Smal, C., and de Prat-Gay, G. (2010). Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral oncoprotein. Structural modularity, intrinsic disorder and phosphorylation of human papillomavirus E7. *FEBS J.* 277, 973–988. doi: 10.1111/j.1742-4658.2009.07540.x

Chen, Z., DeSalle, R., Schiffman, M., Herrero, R., Wood, C. E., Ruiz, J. C., et al. (2018). Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog.* 14:e1007352. doi: 10.1371/journal.ppat.1007352

Cheng, S. J., Chang, C. F., Ko, H. H., Lee, J. J., Chen, H. M., Wang, H. J., et al. (2018). Hypermethylated ZNF582 and PAX1 genes in mouth rinse samples as biomarkers for oral dysplasia and oral cancer detection. *Head Neck* 40, 355–368. doi: 10.1002/hed.24958

Deeks, J. J., and Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *BMJ* 329, 168–169. doi: 10.1136/bmj.329.7458.168

Demarco, M., Hyun, N., Carter-Pokras, O., Raine-Bennett, T. R., Cheung, L., Chen, X., et al. (2020). A study of type-specific HPV natural history and implications for contemporary cervical cancer screening programs. *EClinicalMedicine* 22:100293. doi: 10.1016/j.eclinm.2020.100293

ðukić, A., Lulić, L., Thomas, M., Skelin, J., Bennett Saidu, N. E., Grce, M., et al. (2020). HPV oncoproteins and the ubiquitin proteasome system: a signature of malignancy? *Pathogens* 9:133. doi: 10.3390/pathogens9020133

Durzynska, J., Lesniewicz, K., and Poreba, E. (2017). Human papillomaviruses in epigenetic regulations. *Mutat. Res. Rev. Mutat. Res.* 772, 36–50. doi: 10.1016/j.mrrev.2016.09.006

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Ekanayake Weeramange, C., Tang, K. D., Vasani, S., Langton-Lockton, J., Kenny, L., and Punyadeera, C. (2020). DNA methylation changes in human papillomavirus-driven head and neck cancers. *Cells* 9:E1359. doi: 10.3390/cells9061359

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Forslund, O., Antonsson, A., Nordin, P., Stenquist, B., and Hansson, B. G. (1999). A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *J. Gen. Virol.* 80, 2437–2443. doi: 10.1099/0022-1317-80-9-2437

Huang, J., Wang, G., Tang, J., Zhuang, W., Wang, L. P., Liou, Y. L., et al. (2017). DNA methylation status of PAX1 and ZNF582 in esophageal squamous cell carcinoma. *Int. J. Environ. Res. Public Health* 22:E216. doi: 10.3390/ijerph14020216

Huang, R. L., Chang, C. C., Su, P. H., Chen, Y. C., Liao, Y. P., Wang, H. C., et al. (2012). Methylomic analysis identifies frequent DNA methylation of zinc finger protein 582 (ZNF582) in cervical neoplasms. *PLoS One* 7:e41060. doi: 10.1371/journal.pone.0041060

International Agency for Research on Cancer [IARC] (2012). *Monographs on the Evaluation of Carcinogenic Risks to Humans-Human Papillomaviruses*. Geneva: World Health Organization, 255–313.

Johannessen, L. E., Brandal, P., Myklebust, T. Ã, Heim, S., Micci, F., and Panagopoulos, I. (2018). MGMT gene promoter methylation status - assessment of two pyrosequencing kits and three methylation-specific PCR Methods For Their Predictive Capacity In Glioblastomas. *Cancer Genomics Proteomics* 15, 437–446. doi: 10.21873/cgp.20102

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

Lai, H. C., Lin, Y. W., Huang, T. H., Yan, P., Huang, R. L., Wang, H. C., et al. (2008). Identification of novel DNA methylation markers in cervical cancer. *Int. J. Cancer* 123, 161–167. doi: 10.1002/ijc.23519

Lai, H. C., Ou, Y. C., Chen, T. C., Huang, H. J., Cheng, Y. M., Chen, C. H., et al. (2014). PAX1/SOX1 DNA methylation and cervical neoplasia detection: a Taiwanese Gynecologic Oncology Group (TGOG) study. *Cancer Med.* 3, 1062–1074. doi: 10.1002/cam4.253

Lang, T. A., and Secic, M. (eds). (2006). "Determining the presence or absence of disease: reporting the performance characteristics of diagnostic tests," in *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*, 2nd Edn, (Philadelphia, PA: American College of Physicians), 125–148. Available online at: http://www.acppress-ebooks.com/product/how-to-report-statistics-in-medicine-2nd-edition

Liang, H., Li, G. L., Liu, J., Fu, M., Huang, H., Zhao, K., et al. (2020). The application value of PAX1 and ZNF582 gene methylation in high grade intraepithelial lesion and cervical cancer. *Clin. Transl. Oncol.* [Epub ahead of print]. doi: 10.1007/s12094-020-02416-5

Lin, H., Chen, T. C., Chang, T. C., Cheng, Y. M., Chen, C. H., Chu, T. Y., et al. (2014). Methylated ZNF582 gene as a marker for triage of women with Pap smear reporting low-grade squamous intraepithelial lesions - a Taiwanese Gynecologic Oncology Group (TGOG) study. *Gynecol. Oncol.* 135, 64–68. doi: 10.1016/j.ygyno.2014.08.012

Lioznova, A. V., Khamis, A. M., Artemov, A. V., Besedina, E., Ramensky, V., Bajic, V. B., et al. (2019). CpG traffic lights are markers of regulatory regions in human genome. *BMC Genomics* 20:102. doi: 10.1186/s12864-018-5387-1

Long, J. S., and Freese, J. (eds). (2014). "Models for binary outcomes: interpretation," in *Regression Models for Categorical Dependent Variables Using Stata*, 3rd Edn, (College Station, TX: Stata Press), 227–308.

Martinez-Zapien, D., Ruiz, F. X., Poirson, J., Mitschler, A., Ramirez, J., Forster, A., et al. (2016). Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. *Nature* 529, 541–545. doi: 10.1038/nature16481

Maxouris, C. (2019). *CNN. Today's Google Doodle Honors Dr. George Papanicolaou, a Greek Immigrant Who Invented the Test that Detects Cervical Cancer*. Available online at: https://www.cnn.com/2019/05/13/health/george-papanicolaou-google-doodle-trnd/index.html (Accessed August 1, 2020).

Medvedeva, Y. A., Khamis, A. M., Kulakovskiy, I. V., Ba-Alawi, W., Bhuyan, M. S., and Kawaji, H. (2014). Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* 15:119. doi: 10.1186/1471-2164-15-119

Nakagawa, T., Matsusaka, K., Misawa, K., Ota, S., Takane, K., Fukuyo, M., et al. (2017). Frequent promoter hypermethylation associated with human papillomavirus infection in pharyngeal cancer. *Cancer Lett.* 407, 21–31. doi: 10.1016/j.canlet.2017.08.008

Nayar, R., Chhieng, D. C., Crothers, B., Darragh, T. M., Davey, D. D., Eisenhut, C., et al. (2020). Moving forward-the 2019 ASCCP risk-based management consensus guidelines for abnormal cervical cancer screening tests and cancer precursors and beyond: implications and suggestions for laboratories. *J. Am. Soc. Cytopathol.* 9, 291–303. doi: 10.1016/j.jasc.2020.05.002

Papanicolaou, G. N., and Traut, H. F. (1941). The diagnostic value of vaginal smears in carcinoma of the uterus. *Am. J. Obstet. Gynecol.* 42, 193–206. doi: 10.1016/S0002-9378(16)40621-6

Qiagen (2020). *PyroMark Q48 Autoprep*. Available online at: https://www.qiagen.com/us/products/discovery-and-translational-research/pyrosequencing/instruments/pyromark-q48-autoprep/ (accessed October 1, 2020).

Resnick, R. M., Cornelissen, M. T., Wright, D. K., Eichinger, G. H., Fox, H. S., ter Schegget, J., et al. (1990). Detection and typing of human papillomavirus in archival cervical cancer specimens by DNA amplification with consensus primers. *J. Natl. Cancer Inst.* 82, 1477–1484. doi: 10.1093/jnci/82.18.1477

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Schiffman, M., Clifford, G., and Buonaguro, F. M. (2009). Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect. Agents Cancer* 4:8. doi: 10.1186/1750-9378-4-8

Shen-Gunther, J., Wang, C. M., Poage, G. M., Lin, C. L., Perez, L., Banks, N. A., et al. (2016). Molecular Pap smear: HPV genotype and DNA methylation of ADCY8, CDH8, and ZNF582 as an integrated biomarker for high-grade cervical cytology. *Clin. Epigenet.* 8:96.

Shen-Gunther, J., Wang, Y., Lai, Z., Poage, G. M., Perez, L., and Huang, T. H. (2017). Deep sequencing of HPV E6/E7 genes reveals loss of genotypic diversity and gain of clonal dominance in high-grade intraepithelial lesions of the cervix. *BMC Genomics* 18:231. doi: 10.1186/s12864-017-3612-y

Shen-Gunther, J., and Yu, X. (2011). HPV molecular assays: defining analytical and clinical performance characteristics for cervical cytology specimens. *Gynecol. Oncol.* 123, 263–271. doi: 10.1016/j.ygyno.2011.07.017

Siegel, E. M., Riggs, B. M., Delmas, A. L., Koch, A., Hakam, A., and Brown, K. D. (2015). Quantitative DNA methylation analysis of candidate genes in cervical cancer. *PLoS One* 10:e0122495. doi: 10.1371/journal.pone.0122495

Sigin, V. O., Kalinkin, A. I., Kuznetsova, E. B., Simonova, O. A., Chesnokova, G. G., Litviakov, N. V., et al. (2020). DNA methylation markers panel can improve prediction of response to neoadjuvant chemotherapy in luminal B breast cancer. *Sci. Rep.* 10:9239. doi: 10.1038/s41598-020-66197-1

Sotlar, K., Diemer, D., Dethleffs, A., Hack, Y., Stubner, A., Vollmer, N., et al. (2004). Detection and typing of human papillomavirus by e6 nested multiplex PCR. *J. Clin. Micro* 42, 3176–3184. doi: 10.1128/JCM.42.7.3176-3184.2004

StataCorp (2019). *Stata: Release 16. Multiple Imputation reference manual. Statistical Software*. College Station, TX: StataCorp LLC.

Suarez, I., and Trave, G. (2018). Structural insights in multifunctional papillomavirus oncoproteins. *Viruses* 10:37. doi: 10.3390/v10010037

Tamura, K., Nei, M., and Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11030–11035. doi: 10.1073/pnas.0404206101

Tang, L., Liou, Y. L., Wan, Z. R., Tang, J., Zhou, Y., Zhuang, W., et al. (2019). Aberrant DNA methylation of PAX1, SOX1 and ZNF582 genes as potential biomarkers for esophageal squamous cell carcinoma. *Biomed. Pharmacother.* 120:109488. doi: 10.1016/j.biopha.2019.109488

Tomita, T., Huibregtse, J. M., and Matouschek, A. (2020). A masked initiation region in retinoblastoma protein regulates its proteasomal degradation. *Nat. Commun.* 11:2019. doi: 10.1038/s41467-020-16003-3

Turajlic, S., McGranahan, N., and Swanton, C. (2015). Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochim. Biophys. Acta* 1855, 264–275. doi: 10.1016/j.bbcan.2015.03.005

Ueda, Y., Enomoto, T., Miyatake, T., Ozaki, K., Yoshizaki, T., Kanao, H., et al. (2003). Monoclonal expansion with integration of high-risk type human papillomaviruses is an initial step for cervical carcinogenesis: association of clonal status and human papillomavirus infection with clinical outcome in cervical intraepithelial neoplasia. *Lab. Invest.* 83, 1517–1527. doi: 10.1097/01.LAB.0000092234.68751.83

Ueda, Y., Miyatake, T., Okazawa, M., Kimura, T., Miyake, T., Fujiwara, K., et al. (2008). Clonality and HPV infection analysis of concurrent glandular and squamous lesions and adenosquamous carcinomas of the uterine cervix. *Am. J. Clin. Pathol.* 130, 389–400. doi: 10.1309/ERR93AF840YYNDRQ

van der Zee, R. P., Richel, O., van Noesel, C. J. M., Ciocâf nea-Teodorescu, I., van Splunter, A. P., Ter Braak, T. J., et al. (2020). Cancer risk stratification of anal intraepithelial neoplasia in HIV-positive men by validated methylation markers associated with progression to cancer. *Clin. Infect. Dis.* [Epub ahead of print]. doi: 10.1093/cid/ciaa397

Wentzensen, N., Sherman, M. E., Schiffman, M., and Wang, S. S. (2009). Utility of methylation markers in cervical cancer early detection: appraisal of the state-of-the-science. *Gynecol. Oncol.* 112, 293–299. doi: 10.1016/j.ygyno.2008.10.012

Wild, C. P., Weiderpass, E., and Stewart, B. W. (2020). *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon: International Agency for Research on Cancer.

World Health Organization [WHO] (2014). *Comprehensive Cervical Cancer Control: A Guide to Essential Practice*, 2nd Edn. Geneva: World Health Organization, 23–72.

World Health Organization [WHO] (2020). *WHO Report on Cancer: Setting Priorities, Investing Wisely and Providing Care for All*. Geneva: World Health Organization.

World Health Organization [WHO] News Release (2020). *WHO Outlines Steps to Save 7 Million Lives From Cancer*. Geneva: World Health Organization.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32–35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Conflict of Interest: The United States Army Medical Research and Development Command has filed a patent application on the invention described herein. The inventor is JS-G. No potential conflicts of interest were disclosed by the other authors.

Copyright © 2020 Shen-Gunther, Xia, Stacey and Asusta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 October 2020
doi: 10.3389/fmicb.2020.603509

Differential Expression of Viral Transcripts From Single-Cell RNA Sequencing of Moderate and Severe COVID-19 Patients and Its Implications for Case Severity

*Teng Liu[1], Peilin Jia[1], Bingliang Fang[2,3] and Zhongming Zhao[1,3,4]**

[1] Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [2] Department of Thoracic and Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [3] MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX, United States, [4] Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

OPEN ACCESS

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

Reviewed by:
Li Liu,
Arizona State University, United States
Ping Liang,
Brock University, Canada

***Correspondence:**
Zhongming Zhao
zhongming.zhao@uth.tmc.edu

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

Received: *07 September 2020*
Accepted: *23 September 2020*
Published: *16 October 2020*

Citation:
Liu T, Jia P, Fang B and Zhao Z (2020) Differential Expression of Viral Transcripts From Single-Cell RNA Sequencing of Moderate and Severe COVID-19 Patients and Its Implications for Case Severity. Front. Microbiol. 11:603509. doi: 10.3389/fmicb.2020.603509

With steady increase of new COVID-19 cases around the world, especially in the United States, health care resources in areas with the disease outbreak are quickly exhausted by overwhelming numbers of COVID-19 patients. Therefore, strategies that can effectively and quickly predict the disease progression and stratify patients for appropriate health care arrangements are urgently needed. We explored the features and evolutionary difference of viral gene expression in the SARS-CoV-2 infected cells from the bronchoalveolar lavage fluids of patients with moderate and severe COVID-19 using both single cell and bulk tissue transcriptome data. We found SARS-CoV-2 sequences were detectable in 8 types of immune related cells, including macrophages, T cells, and NK cells. We first reported that the SARS-CoV-2 ORF10 gene was differentially expressed in the severe vs. moderate samples. Specifically, ORF10 was abundantly expressed in infected cells of severe cases, while it was barely detectable in the infected cells of moderate cases. Consequently, the expression ratio of ORF10 to nucleocapsid (N) was significantly higher in severe than moderate cases ($p = 0.0062$). Moreover, we found transcription regulatory sequences (TRSs) of the viral leader sequence-independent fusions with a 5' joint point at position 1073 of SARS-CoV-2 genome were detected mainly in the patients with death outcome, suggesting its potential indication of clinical outcome. Finally, we identified the motifs in TRS of the viral leader sequence-dependent fusion events of SARS-CoV-2 and compared with that in SARS-CoV, suggesting its evolutionary trajectory. These results implicated potential roles and predictive features of viral transcripts in the pathogenesis of COVID-19 moderate and severe patients. Such features and evolutionary patterns require more data to validate in future.

Keywords: COVID-19, SARS-CoV-2, virus evolution, cell infection rate, ORF10, virus gene fusion, viral fusion evolution, virulence

INTRODUCTION

The new coronavirus disease 2019 (COVID-19) has casted an imminent threat to people all over the world. COVID-19 is caused by a novel human-infecting betacoronavirus designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of this manuscript updated writing (September 5, 2020), this life-threatening pneumonia pandemic has quickly caused more than 26.7 million confirmed cases and over 875,000 deaths globally since it first emerged in December 2019 (Sohrabi et al., 2020; Yang et al., 2020). SARS-CoV-2 genome, like SARS-CoV, is under frequent mutations which likely caused the change of infection rate (Zhao et al., 2004). As a positive-sense, single-stranded RNA virus with an approximate 30 kb in length of its genomic RNA (gRNA) (Pal et al., 2020), SARS-CoV-2 infects host cells through cell surface proteins ACE2 and TMPRSS2 (Hoffmann et al., 2020; Zhou et al., 2020). Inside host cells, SARS-CoV-2 gRNA can synthesize subgenomic RNAs (sgRNAs) during its replication (Khailany et al., 2020). The canonical coronaviral sgRNA contains a 5' leader sequence with ∼65 nucleotides (nt) in length (Wu and Brian, 2010), which is fused to the "body" sequence to form the sgRNA with transcription regulatory sequences (TRSs) (Kim et al., 2020). A total of 10 open reading frames have been identified in SARS-CoV-2 gRNA, which generate multiple viral proteins including RNA-dependent RNA polymerase (ORF1ab), spike protein (S), ORF3a, envelope protein (E), membrane protein (M), ORF6, ORF7a, ORF8, nucleocapsid protein (N), and ORF10. Among them, S, E, M, and N are structural proteins that package virus gRNA to generate progeny (Lu et al., 2020; Wolfel et al., 2020). The ORF1ab gene encodes a polyprotein, which plays an important role in viral RNA synthesis. ORF3a, ORF6, ORF7a, ORF8, and ORF10 are accessory proteins. Their functions are not fully characterized (Kim et al., 2020).

Clinically, approximately 80% of laboratory-confirmed cases have moderate symptoms, while the remaining ∼20% have severe disease with dyspnea, acute respiratory distress syndrome and/or multiple organ failure, some of whom require the treatment within intensive care unit (ICU). Prolonged deterioration of lymphopenia and high levels of inflammatory cytokines in plasma were found to be associated with severity of COVID-19 (Liu et al., 2020). Subsets of lymphocytes, including CD8+ T cell, CD4+ T cell, B cell, and NK cell, were all significantly lower in severe COVID-19 patients than the moderate patients (He et al., 2020; Huang et al., 2020; Zheng et al., 2020). A recent study using single-cell RNA sequencing (scRNA-seq) of bronchoalveolar lavage fluid (BALF) samples revealed that BALFs from patients with severe symptoms had more proinflammatory monocyte-derived macrophages, higher levels of inflammatory cytokines (IL8, IL6, and IL1B), and less clonally expanded CD8+ T cells than that from the moderate cases, suggesting that cytokine storm is associated with severity of COVID-19 (Liao et al., 2020). However, molecular mechanisms underlying these different immune microenvironment changes between moderate and severe COVID-19 cases remain uncharacterized. We made use of these scRNA-seq datasets together with bulk RNA-seq datasets obtained from BALFs of COVID-19 patients recently published by others (Liao et al., 2020; Xiong et al., 2020; Zhou et al., 2020) to explore types of host cells infected by SARS-CoV-2 and levels of viral gene expression in infected cells. We observed that SARS-CoV-2 sequences were detected in 8 types of immune related cells. Interestingly, we found viral ORF10 was abundantly expressed in the samples from the severe cases but barely detectable in the samples from the moderate cases. Moreover, TRS leader sequence-independent fusion sequences with a 5' joint point at position 1073 of SARS-CoV-2 genome were detected only in the patients with death outcome. These results implicated that selective expressions of some viral transcripts in infected cells might play roles in severity of COVID-19 and, if so, it might be used as markers for predicting COVID-19 progressions and severity. Our splicing and TRS motif features will help understand how SARS-CoV has evolved at the expression level. We expected more data will be generated in the near future, which will be used to further validate these findings.

MATERIALS AND METHODS

Transcriptome Data

Single-cell RNA sequencing datasets were retrieved from NCBI GEO database (ID: GSE145926[1]). The detailed sequence, demographic (all patients were from Shenzhen, China), and clinical information was provided in **Supplementary Tables S1, S2** in the original study (Liao et al., 2020). In this study, we mainly compared sequence and expression features in the infected vs. uninfected cells from severe cases (n = 6) and moderate cases (n = 3), so the data from normal cases (n = 3) were not used due to lack of infected cells. For all the scRNA-seq samples used in this work, 5′ reagent kits were used for library construction. The bulk RNA-seq datasets were obtained from GEO (ID: GSE150316) and the BIGD database (IDs: CRA002423 and CRA002390[2]).

scRNA-Seq Data Analysis

For analyzing scRNA-seq data, raw reads from scRNA-seq libraries were first cleared by removing the template switch oligo sequence. Reads with adaptor contaminants and low-quality bases were removed. The SARS-CoV-2 reference genome was downloaded from NCBI (MN908947.3, Wuhan-Hu-1 complete genome) including its mapping information of the 10 genes. Alignment of sequence was performed by Cell Ranger software (v. 3.1.0)[3]. Only the reads that were confidently mapped to the SARS-CoV-2 reference genome were used for unique molecular identifier (UMI) counting. We applied UMI > 2 as the cutoff for identification of the infected cells. As a result, distinct UMIs of each gene for a given cell were counted as the transcript copy number of that gene. To reduce the gene expression matrix to its most important features, we used the Uniform Manifold Approximation and Projection (UMAP)

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

[2]https://bigd.big.ac.cn/

[3]https://support.10xgenomics.com/

method (McInnes et al., 2018) implemented in the Seurat R package and then iteratively clustering strategy to perform subsequent cell clustering. Furthermore, we used Seurat (Stuart et al., 2019) to identify the marker genes of cell groups by calling differentially expressed genes (DEGs) between the cells of each sub-group and the remaining cells. DEGs were selected if the fold change of log2-transformed expression level was > 1 and $p < 0.05$ (Xu et al., 2012; Hua et al., 2020).

Bulk RNA-Seq Data Analysis

For analysis of bulk RNA-seq data, raw reads were pre-processed using Cutadapt (v1.15) to remove bases with quality scores < 20 and adapter sequences (Martin, 2011). STAR (v2.7.3a) (Dobin et al., 2013) was used for alignment. Then, Stringtie (2.1.2) was used to normalize the gene expression of SARS-CoV-2. We calculated TPM (Transcripts Per Kilobase Million) for gene expression comparison. All statistical tests in this work were performed based on unpaired *t*-test.

RESULTS

Infected Cells in Bronchoalveolar Lavage Fluids of COVID-19 Patients

To explore host cells infected by SARS-CoV-2 and viral gene expression in the infected cells, we downloaded public scRNA-seq datasets of bronchoalveolar lavage fluid (BALF) samples obtained from SARS-CoV-2 infected patients, including 3 moderate patients and 6 severe patients (**Supplementary Table 1**; Liao et al., 2020). The number of total reads ranged from 570,603,931 to 1,417,728,004. The transcriptomic reads of these 9 samples were aligned to SARS-CoV-2 reference genome to detect infected cells. We used UMI > 2 as cutoff for the infected cells. Using the cell type annotations from Liao et al. (2020), we generated UMAP presentation of major cell types in moderate and severe samples, including information of the infected cell and distribution across these cell types (**Figure 1A**). Our analysis detected 10 cell types. Number of cells varied greatly among those cell types: the top three cell types measured by the largest cell counts were macrophages, T, and epithelia. Among the three moderate cases, only one cell type had measurable infected cells (macrophages) and two cell types had only one infected cell in one or two cases (epithelial and NK). Among the six severe cases, we detected eight cell types with infected cells, though number of infected cells varied widely among those cases (**Supplementary Table 2**). For neutrophils (including both infected and uninfected neutrophils), we detected an average of 267.2 cells in the samples of severe cases, which is 200-fold more than that (an average of 1.3 cells) in the samples of moderate cases. For epithelial cells (including both infected and uninfected epithelial cells), there were 312.3 cells on average in the samples of severe cases vs. 71.7 cells on average in the samples of moderate cases, a 4.36-fold difference (*t*-test, $p = 7.71 \times 10^{-5}$) (**Supplementary Table 2**). The result suggested the presence of severe inflammation and desquamation of lung epithelial cells in the patients with severe COVID-19. We found that the macrophages were highly infiltrated in BALF samples, accounting for 70.2% of all the BALF cells (**Supplementary Table 2**). SARS-CoV-2 reads were detected in macrophages of all the samples, with an average of 1.86% of the infected cells (**Table 1**). Plasma cells were relatively few in BALFs, accounting for 2.36% of all the cells in BALFs, with one severe case (S4) as an exception for having a high number of plasma cells ($n = 814$) in BALFs (**Table 1** and **Supplementary Table 2**). In addition, SARS-CoV-2 reads were detected in natural killer (NK) cells and T cells, suggesting the viral infections in these cells (**Table 1**). We found 1.25–25% NK cells were infected in one moderate sample and three severe samples, respectively (**Table 1**). In addition, 0.4–6.45% of T cells in five of the six severe samples, but none in the moderate samples, had SARS-CoV-2 reads (**Table 1**). Infected lung epithelial cells varied from 0.65 to 8.97% and were detected in all the samples. This result suggested that, in addition to epithelial cells, the immune cells, such as macrophages, plasma cells, T cells, and NK cells, can also be infected by SARS-CoV-2.

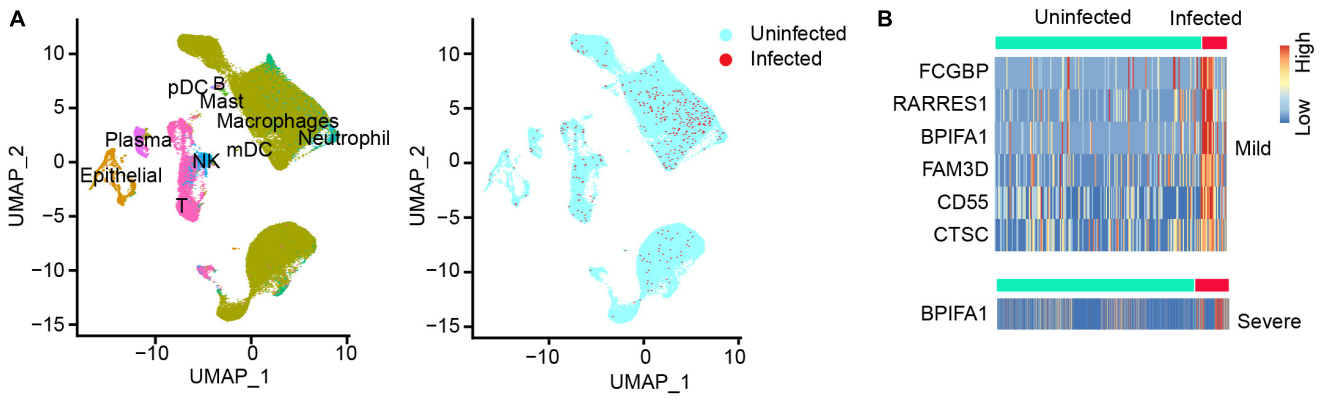


FIGURE 1 | SARS-CoV-2 infected cells and differential gene expression of infected vs. uninfected epithelial cells in BALF samples. **(A)** UMAP (Uniform Manifold Approximation and Projection) view of the cell types for all SARS-CoV-2 infected BALF samples. Color labels each cell cluster identified by scRNA-seq data analysis (see "Materials and Methods"). **(B)** The heatmap of differentially expressed genes (DEGs) of infected epithelial cells vs. uninfected epithelial cells in moderate (mild) cases and in severe cases. One gene, *BPIFA1*, was observed in both moderate and severe cases.

TABLE 1 | Proportion of SARS-CoV-2 infected cells.

Cell type	Moderate (% of total cells)			Severe (% of total cells)					
	M1	M2	M3	S1	S2	S3	S4	S5	S6
B	0	0	0	0	0	0	0	0	0
Epithelial	7.23	1.06	2.63	0.65	6.43	2.16	4.59	3.26	8.97
Macrophages	0.55	0.34	1.72	0.17	0.46	1.41	0.61	0.74	10.71
Mast	0	0	0	0	0	0	0	0	0
mDC	0	0	0	0	0	0	0	0	16.67
Neutrophil	0	0	0	0	0	0	0	0.24	2.33
NK	0	0	3.45	0	0	1.25	2.38	0	25.00
pDC	0	0	0	0	0	0	0	2.38	0
Plasma	0	0	0	0	0	0	2.21	0	0
T	0	0	0	0.04	0	0.16	0.38	0.57	6.45

Differentially Expressed Cellular Genes Between Infected and Uninfected Cells

We called differentially expressed cellular genes in the infected vs. uninfected cells in each cell type. The number of cells in each sub-group is summarized in **Supplementary Table 2**. Using the cutoff of fold change > 2 and $p < 0.05$, we discovered 6 DEGs (*FCGBP, RARRES1, BPIFA1, FAM3D, CD55,* and *CTSC*) which were highly expressed in infected vs. uninfected lung epithelial cells of moderate patients (**Figure 1B**). However, *BPIFA1* was the only gene found highly expressed in the infected vs. uninfected lung epithelial cells of the severe patients (**Figure 1B**). *BPIFA1* encodes a secretory protein and is known to have antimicrobial activity. A recent study showed that *BPIFA1* regulates lung neutrophil infiltration and interferon signaling during acute inflammation (Britto et al., 2019). Its overexpression may play a role in inflammatory response of SARS-CoV-2 infection in lung. Among the other five genes, *FCGBP*, which encodes Fc fragment of IgG binding protein, has been reported with virus infection and viral vector design. *RARRES1* encodes retinoic acid receptor responder 1. It is involved in negative regulation of cell proliferation and endopeptidase activity. *FAM3D*, which encodes FAM3 metabolism regulating signaling molecule D, has Gene Ontology annotations related to cytokine activity. *CD55* encodes a glycoprotein involved in the regulation of the complement cascade. *CTSC* a member of the peptidase C1 family and lysosomal cysteine proteinase. It involves in the activation of many serine proteinases in cells of the immune system.

SARS-CoV-2 Gene Expression in Severe and Moderate Cases

To detect differentially expressed viral genes, we compared the normalized expression of SARS-CoV-2 genes between severe and moderate patients (**Figure 2A**). With the cutoff $p < 0.05$, we found 4 genes (ORF1ab, ORF3a, N, and ORF10) differentially expressed in severe vs. moderate scRNA-seq samples. Interestingly, ORF10 was only detected in the infected cells of severe case samples (**Figure 2B**). All samples from severe cases had a high sequence coverage for ORF10. In contrast, ORF10 was rarely detected in the infected cells of BALF from moderate cases (**Figure 2C**). We compared the ratio of sequence coverage of ORF10 to that of N. This ratio was significantly higher in severe than moderate samples (3.43-folds, $p = 0.0062$, **Figure 2C**). This result suggested that differential expression of ORF10 may be a potential biomarker for COVID-19 progression.

SARS-CoV-2 Genomic Fusions in Host Cells

We observed several viral RNA fusion events (**Supplementary Table 3**) in the scRNA-seq datasets, as well as in bulk RNA-seq datasets from Xiong et al. (2020) and Zhou et al. (2020) (**Supplementary Table 1**). For the transcription regulatory sequence of the viral leader sequence (TRS-L) dependent fusion groups started at position 65, we found a strong signal motif (TCTAAACG) (**Figure 3A** and **Supplementary Table 4**). In addition, we found a highly frequent TRS-L independent fusions that occurred at 5′ position 1073, which fused to multiple 3′ positions without the consistent splicing motif at 3′ locations (**Figure 3B** and **Supplementary Table 5**). TRS-L dependent fusion transcripts are known for the classical sgRNA being formed with a heptameric template-switching signal motif (TCTAAAC), which is critical for viral gene expressions in coronavirus, as reported before COVID-19 outbreak (Wu and Brian, 2010). Here, we reported a very similar, but one nucleotide longer, motif in SARS-CoV-2 genomes (TCTAAACG in SARS-CoV-2 vs. TCTAAAC in SARS-CoV). By manually checking the reads being mapped to TRS-L independent fusions started at 1073 one by one in our data, we found that 14 patients had high-quality reads being mapped to 1,073 (**Table 2**), and 5 out of the 14 patients had the 5' fusion reads at this position. The clinical and sequence information of these 14 patients is provided in **Supplementary Table 1**. Remarkably, all these 5 patients unfortunately passed away (**Table 2**), suggesting that presence of fusions with a 5′ position at 1073 might be associated with worse outcome of COVID-19 patients.

DISCUSSION

Our results revealed that both epithelial cells and immune cells were infected with the SARS-CoV-2. These immune

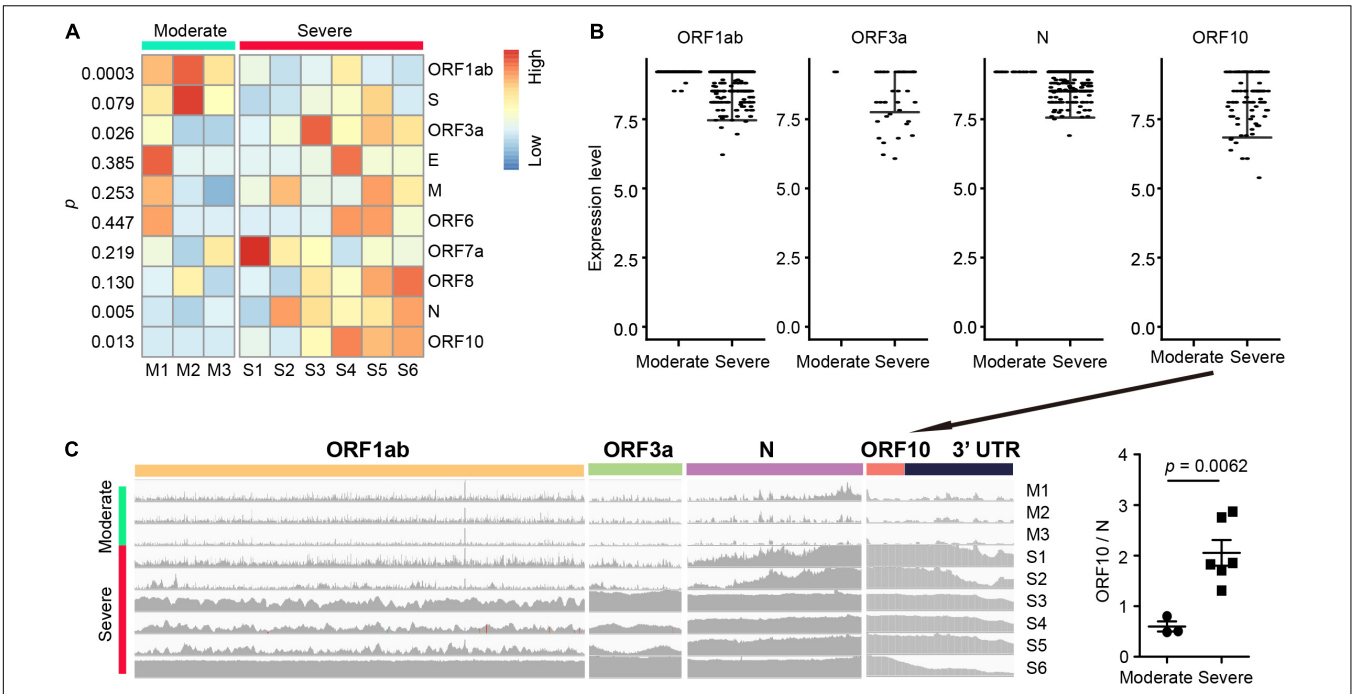


FIGURE 2 | SARS-CoV-2 gene expression and RNA fusions in BALF samples. **(A)** Differential SARS-CoV-2 gene expression (measured by normalized average gene expression) between the moderate and severe groups. **(B)** Expression of four SARS-CoV-2 genes (ORF1ab, ORF3a, N, and ORF10) in each infected cell. Each dot represents an infected cell. **(C)** Levels of read coverages in different viral regions in each sample (left) and ratio of coverage of ORF10 vs. N (right, ORF10/N). The unpaired t test was used for all the statistical tests.

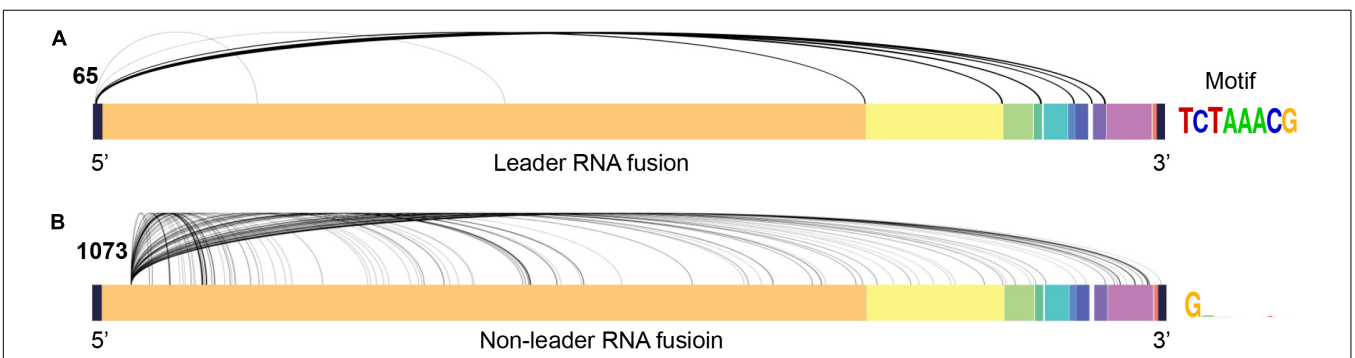


FIGURE 3 | Major viral RNA fusions detected in infected cells. **(A)** TRS-L dependent RNA fusions. **(B)** The most frequent TRS-L independent RNA fusions with a 5' position at 1073 in SARS-CoV-2 genome.

cells included macrophages, monocyte dendritic cells (mDCs), plasmacytoid dendritic cells (pDCs), neutrophils, natural killers (NK), plasma cells, and T cells. The infected epithelial cells and macrophages were found in all moderate and severe cases, indicating that the SARS-CoV-2 virulence could be involved in immune and the related cells. The air way and lung epithelial cells serve important functions as barrier protection for respiratory tract. SARS-CoV-2 is known to infect lung epithelial cells through membrane fusion mediated by binding of SARS-CoV-2's spike to ACE2 on cell surface (Sungnak et al., 2020; Ziegler et al., 2020). Macrophages are the key phagocytes which can take up pathogens. Our results showed that macrophages in all COVID-19 cases were detectable by SARS-CoV-2 sequences. The infected T cells were only found in the severe cases. Middle East respiratory syndrome (MERS) coronavirus can infect T lymphocytes and induce T-cell apoptosis through extrinsic and intrinsic apoptosis pathways (Chu et al., 2016). Whether SARS-CoV-2 also induces T cell apoptosis remains to be determined. *BPIFA1* was highly expressed in both moderate and severe infected epithelial cells, which may play an important role in regulating lung neutrophil infiltration interferon signaling during acute inflammation (Britto et al., 2019).

By comparing SARS-CoV-2 mapped reads between severe and moderate cases, we found that ORF10 was barely expressed in the moderate cases, but it could be detected in high confidence in the severe cases. The ratio of the short-read sequence coverage of

TABLE 2 | Summary of fusion reads starting from SARS-CoV-2 position 1073 in each patient.

Sample name	Type	# Fusion reads	# Total reads	Proportion (%)	Clinical outcome	Source
S3	scRNA-seq	0	255	0	Cured	BALF
S4	scRNA-seq	0	18	0	Cured	BALF
S5	scRNA-seq	1	8	12.50	Death	BALF
S6	scRNA-seq	23	410	7.80	Death	BALF
Case1	Bulk RNA-seq	233	460	50.65	Death	Autopsy
Case2	Bulk RNA-seq	1	1	100	Death	Autopsy
Case5	Bulk RMA-seq	6	6	100	Death	Autopsy
WIV02	Bulk RNA-seq	0	2	0	Cured	BALF
WIV04	Bulk RNA-seq	0	13	0	Cured	BALF
WIV05	Bulk RNA-seq	0	1	0	Cured	BALF
WIV06	Bulk RNA-seq	0	5	0	Cured	BALF
WIV07	Bulk RNA-seq	0	16	0	Cured	BALF
Patient_1	Bulk RNA-seq	0	550	0	Cured	BALF
Patient_2	Bulk RNA-seq	0	193	0	Cured	BALF

The clinical and sequence information of these 14 patients is provided in ***Supplementary Table 1*** *(both scRNA-seq and bulk RNA-seq data).*

ORF10 to that of N in the severe cases was 3.43-fold higher than that in the moderate cases ($p = 0.0062$, **Figure 2C**). The biological significance of the lack of ORF10 expression in infected cells of moderate cases is not yet clear. A recent study on SARS-CoV-2 sgRNAs in viral replication and gene expression in Vero cells suggested that ORF10 might not be expressed as a protein (Kim et al., 2020). Since ORF10 is located close to 3' untranslated region (UTR), it is possible that ORF10 may play a role in regulating transcription or stability of viral sgRNAs. Some previous studies indicated that single-strand virus RNA decays at 3' UTR through nonsense-mediated decay (Balistreri et al., 2014; May et al., 2018). Because only one time point is used for moderate and severe cases in this study, it is not clear whether the difference of ORF10 coverage between moderate and severe cases was related to the decay of SARS-CoV-2's sgRNA and/or gRNA during the disease course. By further exploration of the SARS-CoV-2 reads, we found numerous SARS-CoV-2 RNA genome fusion events in the host cells, suggesting that splicing may be a critical evolutionary mechanism for SARS-CoV-2 to adapt the host cells. Most of them belonged to transcription regulatory sequence of the viral leader sequence (TRS-L) dependent fusion groups, with a strong motif (TCTAAACG). Among TRS-L independent fusions, we found one fusion with a 5' position at 1073. This fusion is most common when compared to other TRS-L independent fusions, and it is only detected in samples of dead cases. Our finding suggested that the presence of these fusions could be associated with worse outcome of COVID-19 patients.

In summary, through re-analysis of the datasets from a recently published scRNA-seq of BALF samples from 9 COVID-19 patients (3 moderate and 6 severe symptoms), we reported that viral ORF10 was highly expressed in severe cases compared to moderate cases. We also reported that the presence of TRS leader sequence-independent fusion sequences with a 5' joint point at position 1073 of SARS-CoV-2 genome was associated with the worst outcome of COVID-19 patients. These results implicated that selective expressions of some viral transcripts might help strengthen virulence If so, such a selective expression might play roles in pathogenesis of COVID-19 and might be used as markers for predicting disease progression and death outcome. Nevertheless, there are several limitations in our results. First, the sample size is very small – we used the scRNA-seq data of bronchoalveolar lavage fluid samples from only 3 moderate patients and 6 severe patients to explore SARS-CoV-2 viral gene expression in the infected cells. Due to the difficulty in obtaining such tissues from COVID-19 patients, this unique dataset is currently what we could find in the field. Second, while the demographic background is consistent (all the samples were collected from Shenzhen, China), age, pre-existing conditions and other clinical factors could not be fully addressed due to small sample size, as discussed in the original study (Liao et al., 2020). The age in the severe case groups was older than that in the moderate case groups, which is consistent with the general clinical reports of older patients having high risk to COVID-19 severity. Third, because the sample size is very small, longitudinal data were not available, and disease progression only covered two stages (moderate and severe; though all the death outcome of the bulk RNA-seq samples in Zhou et al. (2020) and Xiong et al. (2020) studies was confirmed through personal communications.). It remains unclear whether the observed difference was caused by dynamic changes in different time course of SARS-CoV-2 infection in patients. Further validation studies are required before ORF10 expression or presence of 1073 fusions can be used to predict disease progression for COVID-19. As more SARS-CoV-2 genomes will be generated from infected samples in the near future, we will validate our findings when such data are released.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article can be downloaded from NCBI GEO database (IDs: GSE145926 and GSE150316, https://www.ncbi.nlm.nih.gov/geo/) and

the BIGD database (IDs: CRA002423 and CRA002390, https://bigd.big.ac.cn/). Analyzed results will be made available by the authors, without undue reservation, to any qualified researcher. Requests to access these datasets should be directed to GEO and BIGD.

AUTHOR CONTRIBUTIONS

TL and ZZ conceived the study. TL, PJ, BF, and ZZ performed the research and wrote the manuscript. TL collected and analyzed the data. All authors contributed to the article and approved the manuscript.

FUNDING

ZZ was partially supported by the National Institutes of Health grant R01LM012806 and Chair Professorship for Precision Health funds. We thank the technical support from the Cancer Genomics Core funded by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Drs. Yulin Dai and Guangsheng Pei in the Bioinformatics and Systems Medicine Laboratory for technical assistance and valuable discussion. We also thank Zheng-Li Shi and Yu Chen for sharing clinical outcome information.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.603509/full#supplementary-material

REFERENCES

Balistreri, G., Horvath, P., Schweingruber, C., Zund, D., McInerney, G., Merits, A., et al. (2014). The host nonsense-mediated mRNA decay pathway restricts Mammalian RNA virus replication. *Cell Host Microbe* 16, 403–411. doi: 10.1016/j.chom.2014.08.007

Britto, C. J., Niu, N., Khanal, S., Huleihel, L., Herazo-Maya, J. D., Thompson, A., et al. (2019). BPIFA1 regulates lung neutrophil recruitment and interferon signaling during acute inflammation. *Am. J. Physiol. Lung. Cell Mol. Physiol.* 316, L321–L333. doi: 10.1152/ajplung.00056.2018

Chu, H., Zhou, J., Wong, B. H., Li, C., Chan, J. F., Cheng, Z. S., et al. (2016). Middle East respiratory syndrome coronavirus efficiently infects human primary T lymphocytes and activates the extrinsic and intrinsic apoptosis pathways. *J. Infect. Dis.* 213, 904–914. doi: 10.1093/infdis/jiv380

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

He, R., Lu, Z., Zhang, L., Fan, T., Xiong, R., Shen, X., et al. (2020). The clinical course and its correlated immune status in COVID-19 pneumonia. *J. Clin. Virol.* 127:104361. doi: 10.1016/j.jcv.2020.104361

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, 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. *Cell* 181, 271.e8–280.e8. doi: 10.1016/j.cell.2020.02.052

Hua, X., Wang, Y. Y., Jia, P., Xiong, Q., Hu, Y., Chang, Y., et al. (2020). Multi-level transcriptome sequencing identifies COL1A1 as a candidate marker in human heart failure progression. *BMC Med.* 18:2. doi: 10.1186/s12916-019-1469-4

Huang, W., Berube, J., McNamara, M., Saksena, S., Hartman, M., Arshad, T., et al. (2020). Lymphocyte subset counts in COVID-19 patients: a meta-analysis. *Cytometry Part A* 97A, 772–776. doi: 10.1002/cyto.a.24172

Khailany, R. A., Safdar, M., and Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene Rep.* 19:100682. doi: 10.1016/j.genrep.2020.100682

Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell* 181, 914.e10–921.e10. doi: 10.1016/j.cell.2020.04.011

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9

Liu, J., Li, S., Liu, J., Liang, B., Wang, X., Wang, H., et al. (2020). Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine* 55:102763. doi: 10.1016/j.ebiom.2020.102763

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17, 10–12. doi: 10.14806/ej.17.1.200

May, J. P., Yuan, X., Sawicki, E., and Simon, A. E. (2018). RNA virus evasion of nonsense-mediated decay. *PLoS Pathog.* 14:e1007459. doi: 10.1371/journal.ppat.1007459

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3:861. doi: 10.21105/joss.00861

Pal, M., Berhanu, G., Desalegn, C., and Kandi, V. (2020). Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): an update. *Cureus* 12:e7423. doi: 10.7759/cureus.7423

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., et al. (2020). World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* 76, 71–76. doi: 10.1016/j.ijsu.2020.02.034

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888.e21–1902.e21. doi: 10.1016/j.cell.2019.05.031

Sungnak, W., Huang, N., Becavin, C., Berg, M., Queen, R., Litvinukova, M., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* 26, 681–687. doi: 10.1038/s41591-020-0868-6

Wolfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Muller, M. A., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature* 581, 465–469. doi: 10.1038/s41586-020-2196-x

Wu, H. Y., and Brian, D. A. (2010). Subgenomic messenger RNA amplification in coronaviruses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12257–12262. doi: 10.1073/pnas.1000378107

Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., et al. (2020). Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. *Emerg. Microbes Infect.* 9, 761–770. doi: 10.1080/22221751.2020.1747363

Xu, J., Sun, J., Helm, M., Li, A., Chen, J., Zhao, Z., et al. (2012). RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia. *BMC Genomics* 13(Suppl. S8):S2. doi: 10.1186/1471-2164-13-S8-S2

Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan,

China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* 8, 475–481. doi: 10.1016/S2213-2600(20)30079-5

Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y. P., et al. (2004). Moderate mutation rate in the SARS oronavirus genome and its implications. *BMC Evol. Biol.* 4:21. doi: 10.1186/1471-2148-4-21

Zheng, M., Gao, Y., Wang, G., Song, G., Liu, S., Sun, D., et al. (2020). Functional exhaustion of antiviral lymphocytes in COVID-19 patients. *Cell Mol. Immunol.* 17, 533–535. doi: 10.1038/s41423-020-0402-2

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Ziegler, C. G. K., Allon, S. J., Nyquist, S. K., Mbano, I. M., Miao, V. N., Tzouanas, C. N., et al. (2020). SARS-CoV-2 receptor ACE2 Is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* 181, 1016.e19–1035.e19. doi: 10.1016/j.cell.2020.04.035

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.

Copyright © 2020 Liu, Jia, Fang and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 23 October 2020
doi: 10.3389/fmicb.2020.567317

G-Quadruplexes Are Present in Human Coronaviruses Including SARS-CoV-2

Haoran Cui[1,2] *and Leiliang Zhang*[1,2]*

[1]Institute of Basic Medicine, The First Affiliated Hospital of Shandong First Medical University, Jinan, China, [2]Science and Technology Innovation Center, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China

The global coronavirus disease 2019 (COVID-19) pandemic is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is one of seven human coronaviruses. G-quadruplexes are intrinsic obstacles to genome replication. Whether G-quadruplexes are present in human coronaviruses is unknown. In the current study, we have predicted that all seven human coronaviruses harbor G-quadruplex sequences. Conserved G-quadruplex sequences in SARS-CoV and SARS-CoV-2 were analyzed and verified by circular dichroism (CD) spectroscopy and Thioflavin T fluorescence assay. Similar to SARS-CoV, SARS-CoV-2 encodes an nsP3 protein, which is predicted to associate with G-quadruplexes. Targeting G-quadruplex sequences in the SARS-CoV-2 genome by G-quadruplex ligands could be a new way to conquer COVID-19.

Keywords: COVID-19, SARS-CoV-2, SARS-CoV, G-quadruplexes, nsP3

OPEN ACCESS

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

Reviewed by:
Hong-Hui Wang,
Hunan University, China
Chunfu Zheng,
Fujian Medical University, China

***Correspondence:**
Leiliang Zhang
armzhang@hotmail.com

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

Received: *29 May 2020*
Accepted: *23 September 2020*
Published: *23 October 2020*

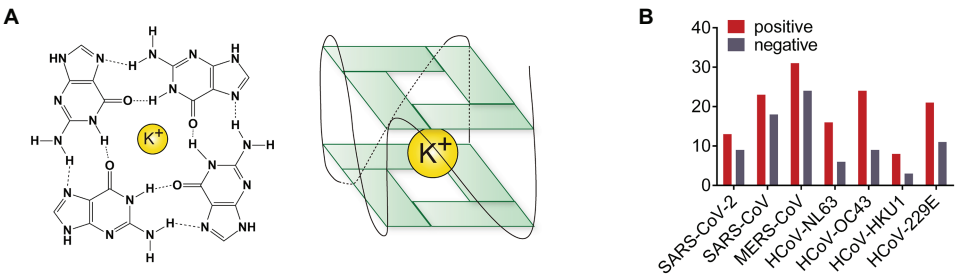
Citation:
Cui H and Zhang L (2020) G-Quadruplexes Are Present in Human Coronaviruses Including SARS-CoV-2. Front. Microbiol. 11:567317. doi: 10.3389/fmicb.2020.567317

INTRODUCTION

Ongoing coronavirus disease 2019 (COVID-19) pandemic has been a major global threat for human health (Lu and Zhang, 2020; Tu et al., 2020), with over 20 million confirmed cases in over 200 countries and regions. COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a betacoronavirus genus of *Coronaviridae* family. Among seven types of *Coronaviridae* family of viruses which could infect humans, HCoV-229E, HCoV-HKU1, HCoV-NL63, and HCoV-OC43 are common around the world, but SARS-CoV, MERS-CoV, and SARS-CoV-2 are more recent and rare.

Both SARS-CoV-2 and SARS-CoV are human SARS-related coronavirus (SARSr-CoVs). SARSr-CoVs have positive-stranded RNA genomes of about 30 kb in length, which encodes multiple proteins. One of the most complex tasks for all viruses is to replicate the entire genome. The replication rate of SARS-CoV-2 is higher than SARS-CoV (Chu et al., 2020). However, the underlying mechanism is not clear. There are intrinsic obstacles to genome replication. For instance, the folding of G-rich sequences into G-quadruplex structures is one source of replication stress. G4 structures' formation requires at least four or more contiguous runs of guanosine nucleotides exist in a short sequence. G-tetrads are formed around K^+ ions through four Hoogsteen-type hydrogen bonds, and then the tetrads stack to adopt G4 structures (**Figure 1A**).

Recently, the important functions for virus G4 structures have been demonstrated. G4 structures in the long terminal repeat promoter of the human immunodeficiency virus (HIV) was critical



C

NCBI accession ID	Sequence name	Sequence length	A% of sequences	T% of sequences	G% of sequences	C% of sequences	GC% of sequences
NC_045512	SARS-CoV-2	29903nt	29.94	32.08	19.61	18.37	37.97
NC_004718	SARS-CoV	29751nt	28.51	30.73	20.80	19.97	40.76
NC_019843	MERS-CoV	30119nt	26.23	32.53	20.93	20.31	41.24
NC_005831	HCoV-NL63	27553nt	26.32	39.22	20.02	14.44	34.64
NC_006213	HCoV-OC43	30741nt	27.66	35.56	21.63	15.16	36.79
NC_006577	HCoV-HKU1	29926nt	27.84	40.10	19.04	13.02	32.06
NC_002645	HCoV-229E	27317nt	27.16	34.58	21.16	16.65	38.26

FIGURE 1 | Human coronaviruses genomic RNA possesses G-quadruplexes. **(A)** Schematic diagram of G-quadruplex structure. Guanines forming G-tetrad structure with a potassium ion inside for stable. **(B)** Numbers of predicted G-quadruplex forming sites in seven human coronaviruses. The red column represents the G-quadruplex numbers in positive strand, and the gray showed the number in negative strand. **(C)** Sequence features of human coronavirus genomes. Lane one to three, NCBI accession ID, names, and sequence length of genome reference sequences. Lane four to seven, the content of each base in genome sequences. Lane eight, GC content in genome sequences.

for promoter activity (Amrane et al., 2014; Perrone et al., 2014). A G4 structure in Epstein-Barr virus (EBV) functioned as a *cis*-acting regulatory region to translate EBV encoded nuclear antigen 1 (EBNA1) mRNA (Murat et al., 2014). G4 structures have also been observed in human papillomavirus (HPV), hepatitis B virus (HBV), Nipah virus, hepatitis C virus (HCV), Zika virus, and Ebola virus (Ruggiero and Richter, 2020). These studies highlight a critical role for G4 structures in viruses.

Whether G4 structures are present in the genomes of human coronaviruses including SARS-CoV-2 is largely unknown. To address this question, we analyzed human coronavirus RNA genomes and predicted several conserved G4. G4 ligands could be developed as antiviral agents for human coronaviruses, including SARS-CoV-2, causing the current COVID-19 pandemic. Moreover, we identified SARS-CoV-2 contains less G4 than SARS-CoV, which partially explains why SARS-CoV-2 replicated faster than SARS-CoV-2.

MATERIALS AND METHODS

Sequence Analysis

Genomes of bat SARSr-CoV, SARS-CoV, and SARS-CoV-2 strains were downloaded from the NCBI virus database. Potential G4 forming sequences in human coronavirus genomes were predicted by QGRS mapper[1] (Kikin et al., 2006) and Quadbase2[2] (Parashar and Shantanu, 2016). The parameters were as below: max. length, 30; min. G-group size, 2; and loop size from 0 to 12. Global genome alignment among human coronavirus was conducted using R package DECIPHER with default parameters (Wright, 2015). The results were exported in Fasta format and visualized in MEGA X. Sequence logos were generated by weblogo tools[3] (Crooks et al., 2004).

Thioflavin T Fluorescence Assay

Single strand RNA oligmers of conserved G4 sequences in SARS-CoV-2, as well as mutant sequences, were synthesized. The sequences were listed in **Table 1**. RNA oligmers were desolved in RNase free buffer containing 20 mM Tris-HCl and 40 mM KCl to a final concentration of 2 μM. Then the RNA solution was heated at 90°C for 5 min and slowly cooling down to room temperature. Thioflavin T (ThT) powder were bought from Aladdin Industrial Corporation and dissolved in the buffer above. Oligmers and ThT were mixed at final concentration 2 and 2 μM, respectively. The fluorescence at 495 nm emission was collected after 425 nm excitation using SpectraMax microplate reader.

Circular Dichroism Spectrum

G4 RNA oligmers were dissolved in RNase free buffer containing 20 mM Tris-HCl and 40 mM KCl to a final concentration of 2 μM. Circular dichroism (CD) spectrum was collected by Chirascan V100 from wavelength 200 to 400 nm.

[1]http://bioinformatics.ramapo.edu/QGRS/index.php/analyze.php
[2]http://quadbase.igib.res.in/TetraPlexFinder
[3]https://weblogo.berkeley.edu/logo.cgi

TABLE 1 | Sequences for ThT staining and CD spectrum.

Name	G-quadruplex or mutant sequence
nsP1-a	5'-UGGCUUUGGAGACUCCGUGGAGGAGGU-3'
nsP1-b	5'-CGGUAAUAAAGGAGCUGGUGGC-3'
nsP10	5'-CGGUAUGUGGAAAGGUUAUGGC-3'
nsP10-m1	5'-CGGUAUGUAGAAAGGUUAUGGC-3'
nsP10-m2	5'-CGGUAUGUAAAAAGGUUAUGGC-3'
S-a	5'-UGGUUGGACCUUUGGUGCAGGU-3'
S-a-m1	5'-UGGUUAGACCUUUGGUGCAGGU-3'
S-a-m2	5'-UGGUUAAACCUUUGGUGCAGGU-3'
S-b	5'-UGGCUUAUAGGUUUAAUGGUAUUGGA-3'
S-c	5'-UGGCCAUGGUACAUUUGGCUAGGU-3'
N	5'-GGGCUGGCAAUGGCGGA-3'
N-m1	5'-GGGCUAGCAAUGGCGGA-3'
N-m2	5'-GGGCUAACAAUGGCGGA-3'
Positive	5'-AGGGCGGUGUGGGAAGAGGGAAGAGGGGGAGGCAG-3'
Negative	5'-GCGCGCGCUUUUGCGCGCGC-3'

Plasmids Construction and Primer Extension Assay

G4 sequences were inserted after start codon of green fluorescent protein (GFP) sequences and cloned into pCHA vector. Primers used for plasmid construction were below: forward primer for nsP10, 5'-CCGGAATTCATGGGTATGTGGAAAGGTTATGGCG TGAGCAAGGGCGCC-3'; forward primer for Sa, 5'-CCGGA ATTCATGGGTTGGACCTTTGGTGCAGGTGTGAGCAAGGG CGCC-3'; forward primer for N, 5'-CCGGAATTCATGGGC TGGCAATGGCGGGTGAGCAAGGGCGCC-3'; and reverse primer for nsP10, Sa, and N, 5'-CGCGGATCCTCACTTGTA CAGCTCATCCAT-3'. Primers used in primer extension assay were below: forward primer, 5'-GTGGAGCAATAGCAGAGCTC-3' and reverse primer, TCACTTGTACAGCTCATCCA-3'. These plasmids were used as PCR templates. N,N′-(9-(4-(Dimethylamino) phenylamino)acridine-3,6-diyl)bis(3-(pyrrolidin-1-yl)propanamide) (BRACO-19) and meso-5,10,15,20-Tetrakis-(N-methyl-4-pyridyl) porphine, Tetratosylate (TMPyP4) at concentrations of 0, 5, 10, 20, and 40 μM were added into PCR reaction mixture. The PCR products were analyzed by electrophoresis in agarose gels stained with Gelred.

Protein Expression Assay

G4-based plasmids expressing GFP constructed above were transfected into 293T cells using PEI transfection reagent. Forty micromolar BRACO-19 and TMPyP4 were added into cells 4 h after transfection and the cell lysis and were collected to detect the expression level of GFP by SDS-PAGE followed by Western blot.

Structure Analysis

The 3D structure of the SARS-CoV-2 nsP3 SARS-unique domain (SUD) was obtained through homology modeling using the swiss-model[4] (Waterhouse et al., 2018). The template was downloaded from the PDB database with PDB ID 2W2G (Tan et al., 2009). The superimposed image and atom distances were generated by chimera software (Downloaded from http://www.rbvi.ucsf.edu/chimera; Pettersen et al., 2004).

[4]https://swissmodel.expasy.org/

G4 Ligands

The previously reported G4 ligands with antiviral function were BRACO-19 (Read et al., 1999), PHENdc3 (De Cian et al., 2007), IZCZ-3 (Hu et al., 2018), PIPER (Fedoroff et al., 1998), PDP (Müller et al., 2010), PDS (Rodriguez et al., 2008), TMPyP4 (Parkinson et al., 2007), c-exNDI (Collie et al., 2012), and quarfloxin (Drygin et al., 2009). The chemical structures were drawn using Chemdraw software.

RESULTS

Prediction of G4 Sequences in Human Coronaviruses

Genome sequences of MERS-CoV, SARS-CoV, SARS-CoV-2, HCoV-NL63, HCoV-229E, HCoV-OC43, and HCoV-HKU1 were obtained from the NCBI nucleotide database. Genomes of bat SARSr-CoV as well as SARS-CoV and SARS-CoV-2 strains were downloaded from the NCBI virus database. Potential G4 forming sequences in human coronavirus genomes were predicted by QGRS mapper (Kikin et al., 2006). In both positive and negative strands, SARS-CoV-2 possessed less number of predicated G4 and less GC content than SARS-CoV (**Figures 1B,C** and **Supplementary Tables S1** and **S2**). SARS-CoV-2 with less G4 sequences replicates faster because of less energy is required to bypass G4 structure.

Global genome alignment among human coronaviruses was conducted and conserved G4 forming sites were analyzed according to the corresponding position in genomes and confirmed manually (**Supplementary Table S3**). Further, another G4 prediction tool quadbase2 was used to confirm the conserved sequences, and the results were listed in **Supplementary Table S4**. As shown in **Supplementary Table S3**, the G4 sequences in ORF1ab were conserved in four coronaviruses, which were SARS-CoV, SARS-CoV-2, HCoV-OC43, and HCoV-229E. The G4 sequences in the S protein-coding region were conserved in five coronaviruses (SARS-CoV, SARS-CoV-2, MERS-CoV, HCoV-NL63, and HCoV-229E), including the three with highest mortality rates. G4 sequences are key RNA secondary structures in the viral genomes. Consequently, these conserved G4 sequences could be interesting targets in developing of innovative drugs against human coronaviruses.

Interestingly, SARS-CoV and SARS-CoV-2 shared the most similar G4 sequences among human coronaviruses. Seven conserved G4 sequences in SARS-CoV and SARS-CoV-2 genomes were listed in **Figure 2A**, and the schematic diagram of G4 sites were shown in **Figure 2B**. There were two in the nsP1 coding region and three in S coding region, and the other two were in nsP10 and N coding regions. According to the SARS-CoV-2 genome annotation, we identified the conserved position of G4 sequences (**Figure 2A**).

Conserved G4 sequences in SARSr-CoV were analyzed, including bat SARS-related coronavirus, SARS-CoV, and SARS-CoV-2. The results showed that the seven G4 sites identified above were also conserved in bat SARSr-CoV (**Supplementary Tables S5** and **S6**). To better visualize the conserved sequences,

sequence logos were generated using the weblogo tool. Results in **Figure 2C** showed that critical G4 sequences were conserved in typical SARSr-CoV. To further explore G4 sequences' evolution in SARS-CoV-2, all the SARS-CoV-2 strains available on the NCBI database (up to 8th April 2020) from different countries were downloaded and analyzed. Genome alignment was conducted using R package DECIPHER, and it turned out that G4 sites were highly conserved in SARS-CoV-2 strains (**Supplementary Table S7** and **Supplementary Figure S1**). Finally, the conserved G4 sequences in all SARSr-CoV strains were aligned and logos were generated (**Supplementary Figure S2**), indicating that the G4 sequences were evolutionarily conserved. Our observation of the strong conservation of the G4 sites in SARSr-CoV genomes supports a hypothesis that these sequences are very important for SARSr-CoV.

Characterization of G-Quadruplex Structures

G4 structure-specific binding to ThT to induce its fluorescence (Renaud de la Faverie et al., 2014). Single strand RNA oligomers of these conserved sequences in SARS-CoV-2 were synthesized. ThT fluorescence assay results showed that these conserved sequences were adopted to form G4 structures in SARS-CoV-2 genomes (**Figure 3A**).

Circular dichroism spectrum was employed to further confirm the existence of G4 structure in SARS-CoV-2 genome. All these seven conserved sequences have absorbance at about 264 nm, indicating that G4 structures in SARS-CoV-2 genome were adopted to form parallel-strand topologies (**Figure 3B**).

To analyze the effect of single guanine in the G4 region, single or two nucleotide mutations of nsP10, S-a, and N G4 sequences were designed and oligomers were synthesized. The results in **Figure 3C** showed that the fluorescence signals of mutation oligomers were decreased significantly compared with that of wide type.

SARS-CoV-2 nsP3 Potentially Associates With G-Quadruplex Sequences

SARSr-CoV encodes an nsP3 protein, which possesses two SUD (M and N) capable of interacting with G4 sequences and potentially essential for unwinding G4 folds in RNA (**Figure 4A**; Kusov et al., 2015). Stimulated structure of SUD from SARS-CoV-2 nsP3 showed a similar structure to SARS-CoV SUD. Interestingly, an L-Y hydrophobic interaction in SARS-CoV-2 SUD replaced the disulfide bond in SARS-CoV SUD (**Figure 4B**). The two SUD from SARS-CoV-2 and SARS-CoV are conserved, indicating the co-evolution of nsP3 and G4 sequences (**Supplementary Figure S3**). The utility of SUD adds additional support for the importance of these genomic G4 folds in SARSr-CoV.

G-Quadruplex Ligands Block G4-Based Gene Expression

G4 ligands could stabilize RNA G4 and have been demonstrated to be potential antivirus strategies for HIV, HBV, HCV, and Ebola virus (Ruggiero and Richter, 2018). **Figure 5A** listed the G4 ligands which have been reported to exert antiviral activities in the past few years. G4 sequences were present in

A

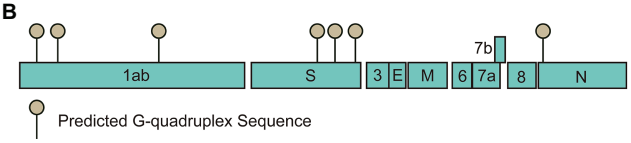
SARS-CoV-2

Position	Length	G-quadruplex sequence	G-Score	Location
353	25	GGCTTTGGAGACTCCGTGGAGGAGG	16	nsP1-a
644	20	GGTAATAAAGGAGCTGGTGG	15	nsP1-b
13385	20	GGTATGTGGAAAGGTTATGG	19	nsP10
24215	20	GGTTGGACCTTTGGTGCAGG	17	S-a
24268	24	GGCTTATAGGTTTAATGGTATTGG	19	S-b
25197	22	GGCCATGGTACATTTGGCTAGG	17	S-c
28903	15	GGCTGGCAATGGCGG	18	N

SARS-CoV

Position	Length	G-quadruplex sequence	G-Score	Location
352	25	GGCTTCGGGGACTCTGTGGAAGAGG	18	nsP1-a
643	20	GGTAATAAGGGAGCCGGTGG	16	nsP1-b
13315	20	GGAATGTGGAAAGGTTATGG	19	nsP10
24090	20	GGATGGACATTTGGTGCTGG	17	S-a
24143	24	GGCATATAGGTTCAATGGCATTGG	19	S-b
25072	22	GGCCTTGGTATGTTTGGCTCGG	17	S-c
28752	15	GGCTAGCGGAGGTGG	17	N

B



C

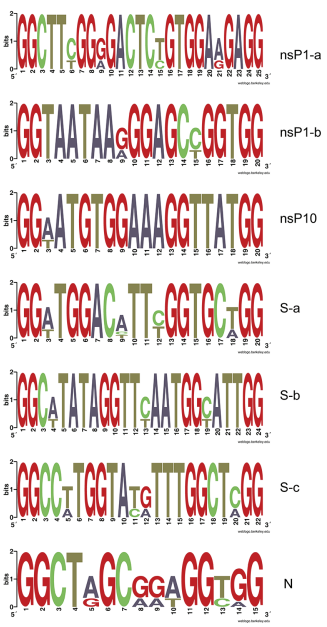


FIGURE 2 | Severe acute respiratory syndrome-related coronavirus (SARSr-CoV) possesses G-quadruplexes. **(A)** Conserved G-quadruplex sequences in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and SARS-CoV genomes. **(B)** A model of the conserved G-quadruplex forming sites in SARS-CoV-2 genome. **(C)** Sequence logos formed by G-quadruplex sequences in typical SARSr-CoV.

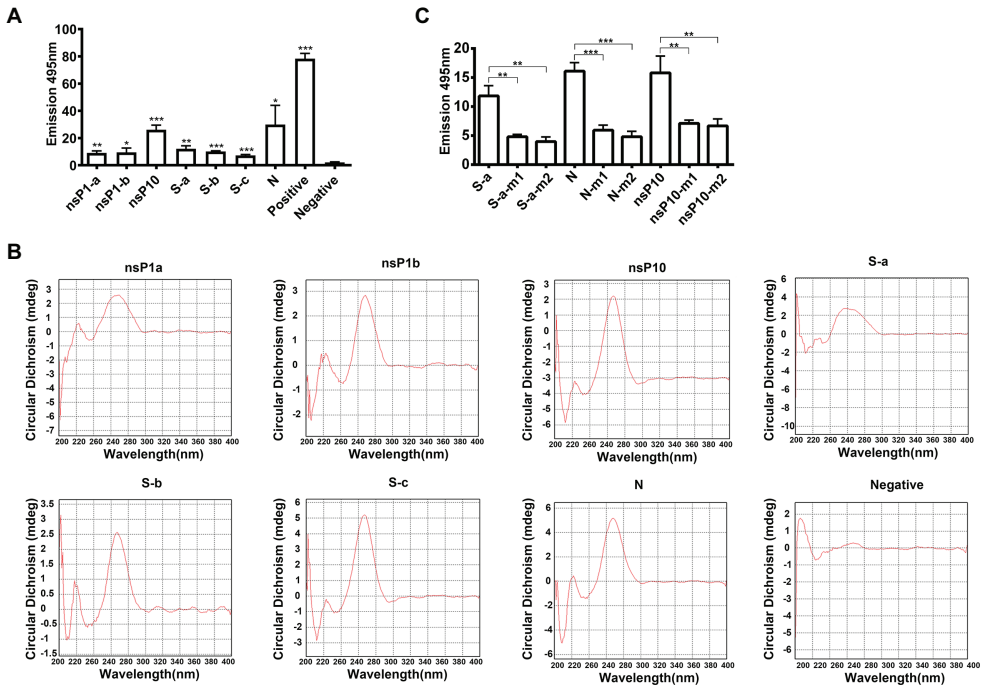


FIGURE 3 | Characterization of G-quadruplex structures. **(A)** Thioflavin T (ThT) fluorescence assay for each conserved G-quadruplex sequences. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$ based on the Student *t*-test. All results are from three independent experiments. Error bars, SD. **(B)** Circular dichroism (CD) spectrum of conserved G-quadruplex sequences. **(C)** ThT fluorescence assay for G-quadruplex single or double mutations. m1, single guanine mutation; m2, double guanines mutation. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$ based on the Student *t*-test. All results are from three independent experiments. Error bars, SD.

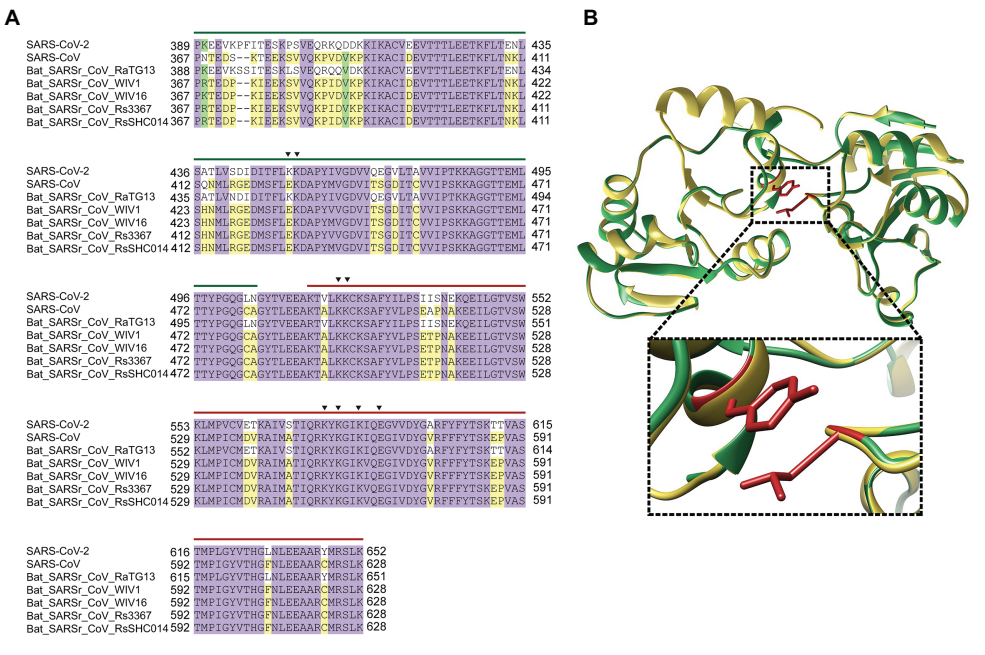


FIGURE 4 | SARS-CoV-2 nsP3 potentially associates with G-quadruplex sequences. **(A)** Sequence alignment of nsP3 SARS-unique domain (SUD)-M and SUD-N in typical SARSr-CoV. The key amino acids in G-quadruplex binding were indicated in triangle. Green line is SUD-M, while red line is SUD-N. **(B)** SUD 3D structure superimpose of SARS-CoV and SARS-CoV-2. The enlarged image showed the atom distance between L516 and Y647 in SARS-CoV-2.

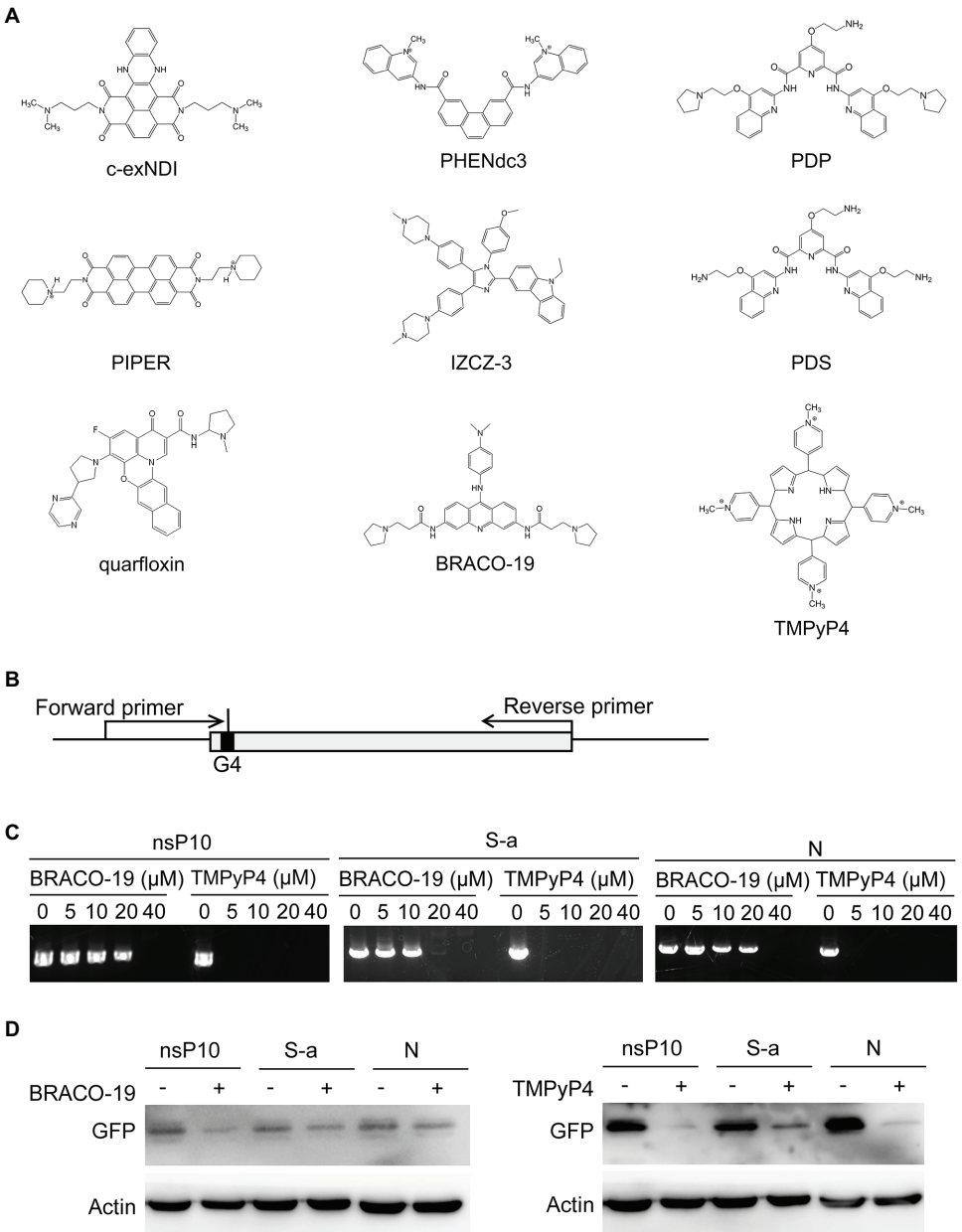


FIGURE 5 | G-quadruplex ligands BRACO-19 and TMPyP4 block G4-based gene expression. **(A)** G-quadruplex ligands with antiviral roles. **(B)** Schematic diagram of G-quadruplex containing plasmids construction and primer extension strategies. **(C)** PCR products of primer extension assay were observed by agarose electrophoresis. The concentrations of G-quadruplex ligands treatment were as indicated. **(D)** GFP expression level with or without treatment were detected through Western blot.

human coronaviruses, and these G4 ligands may be developed as potential drugs against SARS-CoV-2, SARS-CoV, and MERS-CoV.

Primer extension assay were performed to detect whether G4 stabilization affect DNA replication. Plasmids were constructed by inserting the G4 sequences into GFP gene after translation start codon ATG, and two primers were selected to amplify the GFP gene (**Figure 5B**). Two G4 binding ligands, BRACO-19 and TMPyP4, were used to stabilize the G4 structure. As shown in **Figure 5C**, with the increase of treatment concentration, the yield of PCR product reduced. To further detect the influence of G4 structure in protein expression, plasmids constructed above were transfected into cells and GFP expression were examined in the presence or absence of G4 ligands. The results showed both BRACO-19 and TMPyP4 treatment decreased the expression of GFP inserted by G4 sequences (**Figure 5D**).

DISCUSSION

Our study provides a paradigm for assessing G4 functions in viral genomes. Through computational search, different G4-forming sequences were predicted from the human coronaviruses including SARS-CoV-2. The formation of G4s was determined by CD spectroscopy and ThT fluorescence assay. The role of a G4 in gene expression was addressed using primer extension assay and Western blot. Overall, our results point to a potential role for G4s in controlling SARS-CoV-2 viral gene expression. The role of G4s in SARS-CoV-2 viral replication awaits further investigation. We believe that genome-wide analyses of G4s in more viruses will help us to establish a general link between virus life cycle and viral G4s.

G4 structures could hinder gene expression. SARS-CoV-2 contains fewer predicated G4 than SARS-CoV (**Figure 1B**), which partially explains why SARS-CoV-2 replicates faster than SARS-CoV. G4 sequences are potential antiviral targets. We showed G4 ligands including TMPyP4 and BRACO-19 could inhibit G4 reporter expression (**Figure 5D**), indicating that G4 ligands could inhibit G4-contaning virus genome replication.

SARS-CoV-2 G4 could be used to develop tools for SARS-CoV-2 studies. G4 has been applied to detect HCV (Luo et al., 2019) and HIV (Sontakke and Srivatsan, 2020). Likewise, G4 could be developed as a potential biosensor for SARS-CoV-2 detection. G4 is applied to set up HCV helicase assay (Leung et al., 2015). In the future, G4 might be used to measure SARS-CoV-2 nsP13 helicase assay.

Many viral proteins associated with virus-encoded G4. HIV-1 nucleocapsid protein NCp7 binds and unfolds the HIV-1 G4 and promotes reverse transcription (Butovskaya et al., 2019). HCV helicase NS3 unwound viral G4 (Leung et al., 2015). SUD domains of SARS-CoV nsP3 were shown to bind to viral G4 (Tan et al., 2009) and play a critical role in viral replication and transcription (Kusov et al., 2015). Based on the similarity of SARS-CoV SUD and SARS-CoV-2 SUD (**Figure 4B**), nsP3 from SARS-CoV-2 was predicted to associate with viral G4 through SUD domain. Moreover, it is possible that SARS-CoV-2 helicase nsP13 may unwind viral G4 to enhance viral replication.

Viral G4 also associated with host proteins. Cellular nucleolin interacted with viral core G4 to suppress HCV replication (Bian et al., 2019). Nucleolin directly binds to EBV G4 in EBNV1 mRNA sequence to inhibit EBNV1 protein expression (Lista et al., 2017). Nucleolin stabilizes the HIV-1 LTR G4s, and the human ribonucleoprotein A2B1 (HnRNP A2/B1) unwinds the G4s to promote HIV-1 transcription (Scalabrin et al., 2017). Whether host proteins interact with G4 of SARS-CoV-2 remains inconclusive. In the future, we will identify host proteins involved in the function of SARS-CoV-2 G4.

In summary, our results have predicted that all seven human coronaviruses harbor G4 sequences, indicating that G4 structures are crucial elements in the genomes of human coronaviruses. Thus, targeting G4 in viral genomes is a new way to develop antiviral agents. Analysis through genome alignment demonstrated that SARS-CoV and SARS-CoV-2 contained seven conserved G4 sequences. ThT fluorescence assay and CD spectroscopy showed that these conserved G4 sequences in SARS-CoV-2 were able to form G4 folds. Whether these G4s are important for maintaining global genome structure remains an open question. Further studies are needed to better understand these G4s in human coronaviruses and, more specifically, the SARS-CoV-2.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

HC performed the bioinformatics study and experiments. LZ conceived the research. HC and LZ wrote the manuscript and approved the final version for publication.

FUNDING

This work was supported by the National Key Plan for Research and Development of China (2016YFD0500300), National Major S & T Project for the Prevention and Treatment of Major Infectious Diseases in China (2017ZX10004206-007), and Academic Promotion Program of Shandong First Medical University (2019LJ001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.567317/full#supplementary-material

REFERENCES

Amrane, S., Kerkour, A., Bedrat, A., Vialet, B., Andreola, M. L., and Mergny, J. L. (2014). Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. *J. Am. Chem. Soc.* 136, 5249–5252. doi: 10.1021/ja501500c

Bian, W. X., Xie, Y., Wang, X. N., Xu, G. H., Fu, B. S., Li, S., et al. (2019). Binding of cellular nucleolin with the viral core RNA G-quadruplex structure suppresses HCV replication. *Nucleic Acids Res.* 47, 56–68. doi: 10.1093/nar/gky1177

Butovskaya, E., Soldà, P., Scalabrin, M., Nadai, M., and Richter, S. N. (2019). HIV-1 nucleocapsid protein unfolds stable RNA G-quadruplexes in the viral genome and is inhibited by G-quadruplex ligands. *ACS Infect. Dis.* 5, 2127–2135. doi: 10.1021/acsinfecdis.9b00272

Chu, H., Chan, J. F., Wang, Y., Yuen, T. T., Chai, Y., Hou, Y., et al. (2020). Comparative replication and immune activation profiles of SARS-CoV-2 and SARS-CoV in human lungs: an ex vivo study with implications for the pathogenesis of COVID-19. *Clin. Infect. Dis.* 71, 1400–1409. doi: 10.1093/cid/ciaa410

Collie, G. W., Promontorio, R., Hampel, S. M., Micco, M., Neidle, S., and Parkinson, G. N. (2012). Structural basis for telomeric G-quadruplex targeting by naphthalene diimide ligands. *J. Am. Chem. Soc.* 134, 2723–2731. doi: 10.1021/ja2102423

Crooks, G. E., Hon, G., Chandonia, J. -M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

De Cian, A., Cristofari, G., Reichenbach, P., De Lemos, E., Monchaud, D., Teulade-Fichou, M. P., et al. (2007). Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17347–17352. doi: 10.1073/pnas.0707365104

Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., et al. (2009). Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. *Cancer Res.* 69, 7653–7661. doi: 10.1158/0008-5472.can-09-1304

Fedoroff, O. Y., Salazar, M., Han, H., Chemeris, V. V., Kerwin, S. M., and Hurley, L. H. (1998). NMR-based model of a telomerase-inhibiting compound bound to G-quadruplex DNA. *Biochemistry* 37, 12367–12374. doi: 10.1021/bi981330n

Hu, M. H., Wang, Y. Q., Yu, Z. Y., Hu, L. N., and Tan, J. H. (2018). Discovery of a new four-leaf clover-like ligand as a potent c-MYC transcription inhibitor specifically targeting the promoter G-quadruplex. *J. Med. Chem.* 61, 2447–2459. doi: 10.1021/acs.jmedchem.7b01697

Kikin, O., D'Antonio, L., and Bagga, P. S. (2006). QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* 34, W676–W682. doi: 10.1093/nar/gkl253

Kusov, Y., Tan, J., Alvarez, E., Enjuanes, L., and Hilgenfeld, R. (2015). A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication–transcription complex. *Virology* 484, 313–322. doi: 10.1016/j.virol.2015.06.016

Leung, K. H., He, H. Z., He, B., Zhong, H. J., Lin, S., Wang, Y. T., et al. (2015). Label-free luminescence switch-on detection of hepatitis C virus NS3 helicase activity using a G-quadruplex-selective probe. *Chem. Sci.* 6, 2166–2171. doi: 10.1039/c4sc03319a

Lista, M. J., Martins, R. P., Billant, O., Contesse, M. A., Findakly, S., Pochard, P., et al. (2017). Nucleolin directly mediates Epstein-Barr virus immune evasion through binding to G-quadruplexes of EBNA1 mRNA. *Nat. Commun.* 8:16043. doi: 10.1038/ncomms16043

Lu, Y., and Zhang, L. (2020). Social media WeChat infers the development trend of COVID-19. *J. Infect.* 81, e82–e83. doi: 10.1016/j.jinf.2020.03.050

Luo, X., Xue, B., Feng, G., Zhang, J., Lin, B., Zeng, P., et al. (2019). Lighting up the native viral RNA genome with a fluorogenic probe for the live-cell visualization of virus infection. *J. Am. Chem. Soc.* 141, 5182–5191. doi: 10.1021/jacs.8b10265

Müller, S., Kumari, S., Rodriguez, R., and Balasubramanian, S. (2010). Small-molecule-mediated G-quadruplex isolation from human cells. *Nat. Chem.* 2, 1095–1098. doi: 10.1038/nchem.842

Murat, P., Zhong, J., Lekieffre, L., Cowieson, N. P., Clancy, J. L., Preiss, T., et al. (2014). G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.* 10, 358–364. doi: 10.1038/nchembio.1479

Parashar, D., and Shantanu, C. (2016). QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.* 44, W277–W283. doi: 10.1093/nar/gkw425

Parkinson, G. N., Ghosh, R., and Neidle, S. (2007). Structural basis for binding of porphyrin to human telomeres. *Biochemistry* 46, 2390–2397. doi: 10.1021/bi062244n

Perrone, R., Butovskaya, E., Daelemans, D., Palù, G., Pannecouque, C., and Richter, S. N. (2014). Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19. *J. Antimicrob. Chemother.* 69, 3248–3258. doi: 10.1093/jac/dku280

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., and Meng, E. C. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Read, M. A., Wood, A. A., Harrison, J. R., Gowan, S. M., Kelland, L. R., Dosanjh, H. S., et al. (1999). Molecular modeling studies on G-quadruplex complexes of telomerase inhibitors: structure-activity relationships. *J. Med. Chem.* 42, 4538–4546. doi: 10.1021/jm990287e

Renaud de la Faverie, A., Guédin, A., Bedrat, A., Yatsunyk, L. A., and Mergny, J. L. (2014). Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic Acids Res.* 42:e65. doi: 10.1093/nar/gku111

Rodriguez, R., Müller, S., Yeoman, J. A., Trentesaux, C., Riou, J. F., and Balasubramanian, S. (2008). A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J. Am. Chem. Soc.* 130, 15758–15759. doi: 10.1021/ja805615w

Ruggiero, E., and Richter, S. N. (2018). G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. *Nucleic Acids Res.* 46, 3270–3283. doi: 10.1093/nar/gky187

Ruggiero, E., and Richter, S. N. (2020). Viral G-quadruplexes: new frontiers in virus pathogenesis and antiviral therapy. *Annu. Rep. Med. Chem.* 54, 101–131. doi: 10.1016/bs.armc.2020.04.001

Scalabrin, M., Frasson, I., Ruggiero, E., Perrone, R., Tosoni, E., Lago, S., et al. (2017). The cellular protein hnRNP A2/B1 enhances HIV-1 transcription by unfolding LTR promoter G-quadruplexes. *Sci. Rep.* 7:45244. doi: 10.1038/srep45244

Sontakke, V. A., and Srivatsan, S. G. (2020). A dual-app nucleoside probe reports G-quadruplex formation and ligand binding in the long terminal repeat of HIV-1 proviral genome. *Bioorg. Med. Chem. Lett.* 30:127345. doi: 10.1016/j.bmcl.2020.127345

Tan, J., Vonrhein, C., Smart, O. S., Bricogne, G., Bollati, M., Kusov, Y., et al. (2009). The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog.* 5:e1000428. doi: 10.1371/journal.ppat.1000428

Tu, H., Tu, S., Gao, S., Shao, A., and Sheng, J. (2020). The epidemiological and clinical features of COVID-19 and lessons from this global infectious public health event. *J. Inf. Secur.* 81, 1–9. doi: 10.1016/j.jinf.2020.04.011

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427

Wright, E. S. (2015). DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* 16:322. doi: 10.1186/s12859-015-0749-z

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.

Copyright © 2020 Cui and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cellular and Infection Microbiology

REVIEW
published: 27 October 2020
doi: 10.3389/fcimb.2020.587933

Host-Malaria Parasite Interactions and Impacts on Mutual Evolution

Xin-zhuan Su[1]*, *Cui Zhang*[1] and *Deirdre A. Joy*[2]*

[1] Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States, [2] Parasitology and International Programs Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States

Malaria is the most deadly parasitic disease, affecting hundreds of millions of people worldwide. Malaria parasites have been associated with their hosts for millions of years. During the long history of host-parasite co-evolution, both parasites and hosts have applied pressure on each other through complex host-parasite molecular interactions. Whereas the hosts activate various immune mechanisms to remove parasites during an infection, the parasites attempt to evade host immunity by diversifying their genome and switching expression of targets of the host immune system. Human intervention to control the disease such as antimalarial drugs and vaccination can greatly alter parasite population dynamics and evolution, particularly the massive applications of antimalarial drugs in recent human history. Vaccination is likely the best method to prevent the disease; however, a partially protective vaccine may have unwanted consequences that require further investigation. Studies of host-parasite interactions and co-evolution will provide important information for designing safe and effective vaccines and for preventing drug resistance. In this essay, we will discuss some interesting molecules involved in host-parasite interactions, including important parasite antigens. We also discuss subjects relevant to drug and vaccine development and some approaches for studying host-parasite interactions.

Keywords: ***Plasmodium*, genome diversity, population, vaccine, immunity, selection**

OPEN ACCESS

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

Reviewed by:
Fang Huang,
National Institute of Parasitic Diseases, China
Liwang Cui,
University of South Florida, United States

***Correspondence:**
Xin-zhuan Su
xsu@niaid.nih.gov
Deirdre A. Joy
djoy@niaid.nih.gov

Specialty section:
This article was submitted to Parasite and Host, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *11 August 2020*
Accepted: *22 September 2020*
Published: *27 October 2020*

Citation:
Su X-z, Zhang C and Joy DA (2020) Host-Malaria Parasite Interactions and Impacts on Mutual Evolution. Front. Cell. Infect. Microbiol. 10:587933. doi: 10.3389/fcimb.2020.587933

INTRODUCTION

Malaria Parasites, Life Cycle, and Genome

Malaria is a parasitic disease caused by *Plasmodium* species (spp.), unicellular protozoan organisms in the phylum of *Apicomplexa*. The species that infect humans include *Plasmodium falciparum, P. vivax, P. malariae, P. knowlesi,* and *P. ovale*, with *P. ovale* recently recognized as two subspecies called *Plasmodium ovale curtisi* (classic type) and *Plasmodium ovale wallikeri* (Sutherland et al., 2010). Whereas, *P. vivax* is the most widespread species, *P. falciparum* is the most deadly to humans. There are several other species of malaria parasites that infect non-human primates including multiple species that infect African apes, and others that infect Asian and New World monkeys (Escalante et al., 2005; Liu et al., 2010; Boundenga et al., 2015; Martinelli and Culleton, 2018; Galinski, 2019). There are also malaria parasites of rodents such as *P. berghei, P. chabaudi,* and *P. yoelii*, often used as animal models of malaria, as well as parasites that infect birds (*P. gallinaceum* and *P. relictum*) (Duval and Ariey, 2012; Ramiro et al., 2012; Fecchio et al., 2020), lizards, bats, and ungulates (Schall, 1982; Schaer et al., 2013; Templeton et al., 2016; Galen et al., 2018). Human malaria cases are mostly reported in tropical and subtropical regions, particularly in Africa, south

and southeast Asia, and south and central America. In 2018, an estimated 228 million cases of malaria occurred worldwide, with ∼213 million (or 93%) in Africa (WHO, 2019).

Human malaria infections start with a bite from an infected female *Anopheles* mosquito, injecting sporozoites into the skin of a host (**Figure 1**). The sporozoites move to the liver through blood vessels and then traverse liver sinusoidal endothelial cells or Kupffer cells to infect hepatocytes (Baer et al., 2007; Tavares et al., 2013). Inside a hepatocyte, a parasite divides into thousands of merozoites through a process called schizogony. Mature merozoites enter the bloodstream after rupture of the infected hepatocyte and invade erythrocytes, starting a new cycles of schizogony within red blood cells (RBCs) which includes asexual replication of their haploid genome. Within RBCs parasites develop through ring, trophozoite, and schizont stages. The resulting mature schizont is segmented in appearance and contains 16–32 daughter merozoites. The infected RBC (iRBC) ruptures, releasing the daughter sporozoites to invade new RBCs. This cycle of erythrocytic development takes approximately 24 h for *P. knowlesi*, 48 h for *P. falciparum, P. vivax*, and *P. ovale*, and 72 h for *P. malariae*. In addition to merozoites, ruptured iRBCs also release various by-products of parasite metabolism such as hemozoin formed when the parasite digests hemoglobin. The release of parasite materials also triggers host responses leading to various clinical symptoms, including fever, chills, headache, dizziness, back pain, and myalgia (Ashley et al., 2018). Some patients may progress to severe malaria with coma (cerebral malaria), pulmonary edema, acute renal failure, jaundice, severe anemia, acidosis, hypoglycemia, and death (Trampuz et al., 2003; Sypniewska et al., 2017; Luzolo and Ngoyi, 2019). In response to changes in the host environment such as variation in metabolites and immune factors (Josling and Llinas, 2015; Brancucci et al., 2017), some of the parasites differentiate into sexual stages termed male and female gametocytes. When a mosquito takes a bloodmeal from an infected host, gametocytes will undergo sexual development in the midgut of the mosquito. Within minutes of entering the mosquito midgut, male and female gametocytes differentiate into male and female gametes that fertilize to form zygotes and then into motile ookinetes. Ookinetes penetrate through the mosquito midgut wall to develop into oocysts containing thousands of sporozoites. Mature sporozoites migrate to salivary glands and are injected into a new vertebrate host when the mosquito bites again, completing the life cycle.

The genomes of many malaria parasites infecting humans (*P. falciparum, P. vivax, P. ovale, P. malariae, P. knowlesi*), non-human primates (*P. reichenowi, P. praefalciparum, P. blacklocki, P. adleri, P. billcollinsi, P. gaboni, P. cynomolgi, P. coatneyi*, *P. inui, and P. fragile*), rodents (*P. berghei, P. yoelii, P. chabaudi,* and *P. vinkei*) and birds (*P. relictum* and *P. gallinaceum*) have been sequenced and analyzed (Shutler et al., 2005; Pain et al., 2008; Otto et al., 2014a,b, 2018; Lauron et al., 2015; Ansari et al., 2016; Auburn et al., 2016; Sundararaman et al., 2016; Pasini et al., 2017; Rutledge et al., 2017; Bohme et al., 2018; Su et al., 2019). Malaria parasites have relatively small haploid genomes, ranging from 20 to 35 megabases (Mb) that contain 14 chromosomes, a circular plastid genome of ∼35 kb, and multiple copies of a 6-kb mitochondrial DNA (Gardner et al., 2002; Otto et al., 2014a,b; Auburn et al., 2016; Pasini et al., 2017; Bohme et al., 2018). The homologous genes in different species of malaria parasites are often found in syntenic blocks arranged in different order on the chromosomes (Carlton et al., 2005; Kooij et al., 2005). For more information on *Plasmodium* genome and genomics associated with parasite development, diseases, diagnosis, epidemiology, and vaccine development please consult (Kirchner et al., 2016; Bourgard et al., 2018; Escalante and Pacheco, 2019; Galinski, 2019; Garrido-Cardenas et al., 2019; Su et al., 2019; Videvall, 2019).

EVOLUTION AND ADAPTATION OF MALARIA PARASITES IN GREAT APES

Origin of *P. falciparum*

Our understanding of the host-parasite interactions involved in the emergence of *P. falciparum* in humans has been greatly enabled by the discovery of multiple closely related malaria species in African apes beginning in the early 2000s. Three malaria species morphologically similar to *P. falciparum* were described in African apes during the last century (Reichenow, 1920; Blacklock and Adler, 1922; Adler, 1923), and of these only one, *P. reichenowi*, had been sequenced (Neafsey et al., 2005). This picture began to change with the discovery of a new malaria species identified in chimpanzees kept as pets in a village in Gabon (Ollomo et al., 2009). Based on the complete mitochondrial genome this parasite was found to fall within the *Laverania* subgenus that includes *P. reichenowi* and *P. falciparum*. A number of other studies quickly followed, leading to the discovery of additional novel malaria species, and a rethinking of the origin of *P. falciparum* in humans. DNA sequences of the cytochrome oxidase b (*cyt b*) gene in samples from a combination of both wild-living and captive chimpanzees and gorillas showed an unexpectedly diverse collection of *Plasmodium* lineages circulating in African apes, suggesting that *P. falciparum* likely originated from *P. reichenowi* via a single transfer event from chimpanzees to humans (Prugnolle et al., 2010). Sequencing of the *Plasmodium* mitochondrial genome and two nuclear genes, dihydrofolate reductase-thymidylate synthase (*dhfr-ts*) and the gene encoding the merozoite surface protein 2 (*msp2*), from blood samples of captive chimpanzees and bonobos in Uganda and the Democratic Republic of Congo identified two new *Plasmodium* species, and led the authors to conclude that *P. falciparum* arose in bonobos (*Pan paniscus*) (Krief et al., 2010). Yet another study of a small number of chimpanzees and gorillas from Cameroon previously kept as pets targeted two mitochondrial genes (*cytb* and *cox1*), one plastid gene (*tufA*), and one nuclear gene (*ldh*), and purportedly found *P. falciparum* in both chimpanzee and gorilla blood samples (Duval et al., 2010).

The contradictory findings on the origin of *P. falciparum* could be explained in part by sampling limitations, including small sample sizes and a general reliance on captive animals which introduces the possibility of contemporary infection by humans. These concerns were addressed in a large-scale study of non-invasive samples from wild-living chimpanzees, gorillas,

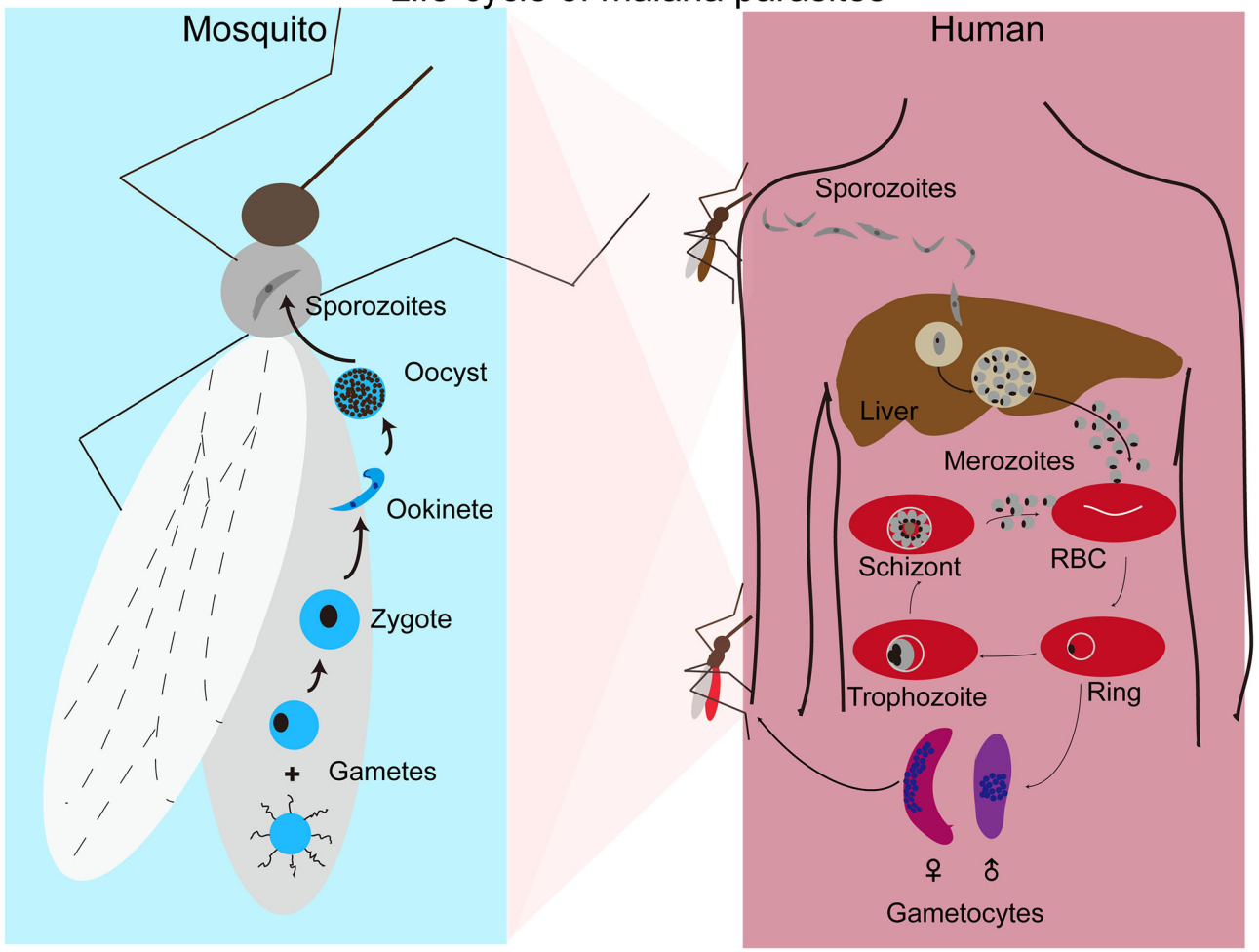


FIGURE 1 | Life cycle of malaria parasites. Malaria parasite infection starts with a mosquito bite injecting sporozoites into human body where the parasites travel to liver and invade hepatocytes. After replication in liver cells, thousands of mature merozoites are released into blood stream and invade red blood cells (RBCs). The parasites then replicate in RBCs that subsequently rupture, releasing more merozoites to invade new RBCs for another cycle. Malaria symptoms correlate with the rupture of infected RBCs. A small number of merozoites develop into male and female gametocytes after invading new RBCs. Gametocytes differentiate into male and female gametes in the mosquito midgut when another mosquito takes a blood meal. Male and female gametes fertilize to produce zygotes that develop into motile ookinetes and oocysts after penetrating the mosquito midgut. A large number of sporozoites from each oocysts moves to salivary glands and are injected into a new human host when the mosquito bites again.

and bonobos throughout central Africa (Liu et al., 2010). Mitochondrial, apicoplast, and nuclear genes from over 1,000 samples were sequenced, and most infected samples were found to contain multiple genetically diverse parasites, highlighting another potential problem with the previous studies. In samples with multiple genetically diverse parasites, conventional PCR methods risk generating recombinant artifacts that could mislead the data. To guard against this, samples were subjected to limiting dilution such that each positive PCR reaction contained no more than a single parasite. Using this approach, the number of host-specific parasite species within the *Laverania* subgenus was increased to seven: *P. gaboni*, *P. billcollinsi,* and *P reichenowi* in chimpanzees; *P. praefalciparum*, *P. blackloci,* and *P. adleri* in gorillas; and *P. falciparum* in humans (Liu et al., 2010). One additional parasite species, *Plasmodium lomamiensis*, was later identified in bonobos (Liu et al., 2017). *P. praefalciparum* in western gorillas was found to be genetically nearly identical to *P. falciparum*, which formed a monophyletic clade (**Figure 2**), strongly suggesting that the *P. falciparum* derived from a host switch from gorillas to humans. Genome sequences from *P. reichenowi* and two *P. gaboni* parasites would later allow for an estimation of the genetic diversity of these two species (Sundararaman et al., 2016). Both *P. reichenowi* and *P. gaboni* were estimated to be at least 10X more diverse than *P. falciparum,* strongly suggesting a relatively recent origin for *P. falciparum*. This was confirmed when additional *Laverania* genomes were used to estimate the divergence of *P. falciparum* and *P. praefalciparum* to be 40,000–60,000 years ago (Otto et al., 2018).

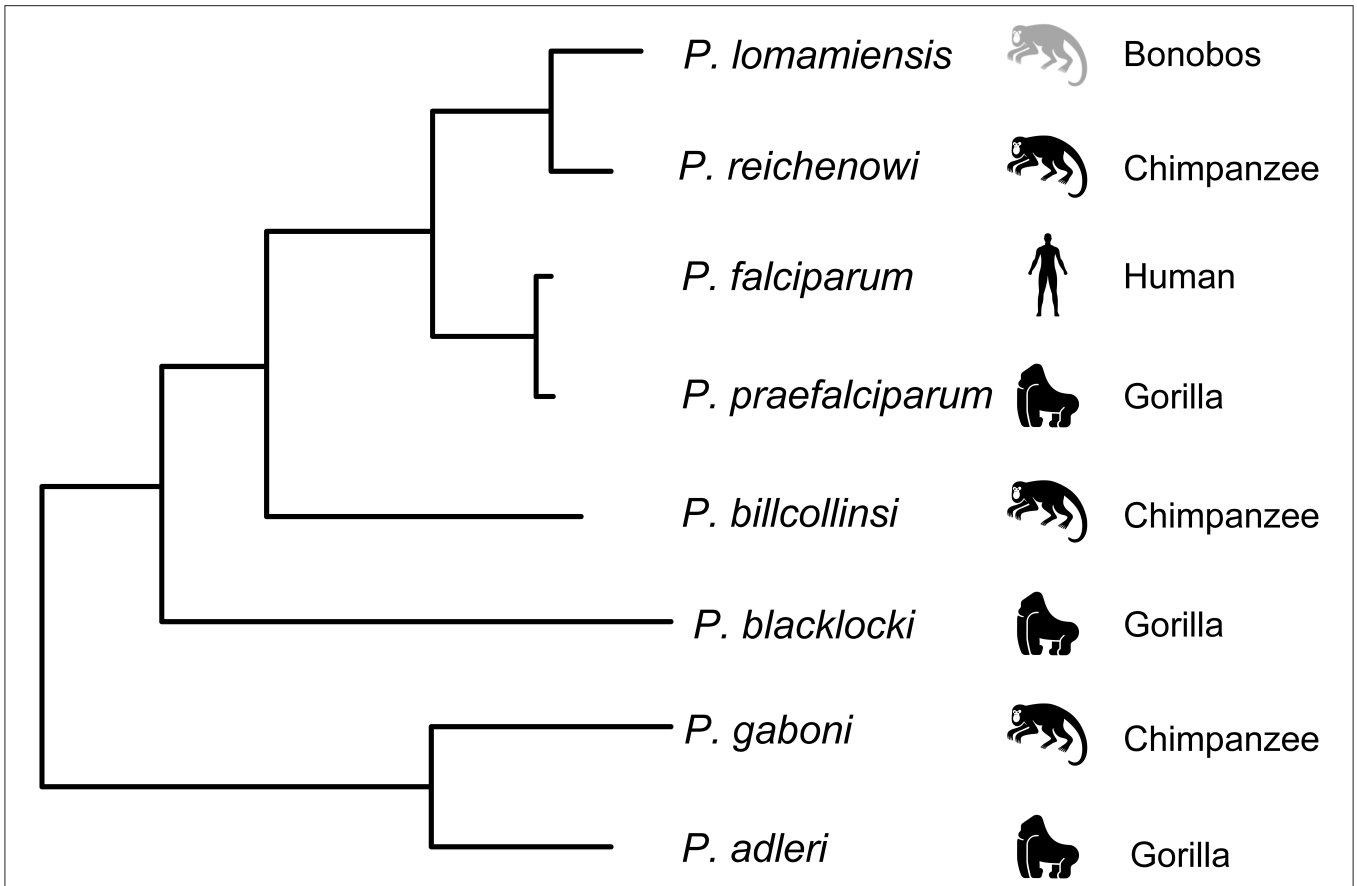


FIGURE 2 | Relationship among the *Laverania* species, adapted from Figure 1A in Plenderleith et al. (2019). Icons indicate the parasites usual host species: gorillas (*Gorilla gorilla*), chimpanzees (*Pan troglodytes*), bonobos (*Pan paniscus*), and humans. The tree was derived by maximum likelihood analysis of a 3.4 kb region of the mitochondrial genome (see Supplementary Figure 1A in Liu et al., 2010).

Additionally, the successful jump to humans by *P. falciparum* appears to have involved multiple parasites over time rather than a single event (Otto et al., 2018). The relationships of *Laverania* species based on maximum likelihood analysis of a 3.4 kb region of the mitochondrial genome are presented in **Figure 2**.

Molecular Basis of Host Trophism and Adaptation to Humans

The *Laverania* exhibit strong host trophism, with each parasite species infecting a single host species, although rare host switching events have occurred. Host trophism is not seen to the same extent in malaria parasites that infect Asian primates (*P. cynomolgi* and *P. knowlesi*). Barriers restricting parasite species to a single host species could exist at any of the host-parasite-vector interfaces, although considerable attention has focused on host-parasite interactions during the blood stage, specifically erythrocyte invasion (Martin et al., 2005; Rich et al., 2009; Rayner et al., 2011; Wanaguru et al., 2013; Galaway et al., 2019; Plenderleith et al., 2019; Proto et al., 2019).

The production of a greatly improved *P. reichenowi* reference genome, as well as a partial genome of *P. gaboni*, allowed for the use of comparative genomics to look for genetic loci involved in host restriction. The genomes of *P. reichenowi, P. gaboni,* and *P. falciparum* showed good conservation and synteny, but with striking differences in some erythrocyte invasion loci (Otto et al., 2014b). A large-scale genomics study of *Laverania* including multiple genomes from all known species found evidence for three interspecific gene transfers (Otto et al., 2018), one which would later be challenged by a reanalysis of the sequence reads (Plenderleith et al., 2019). Several genes involved in either invasion or pathogenesis were fixed in parasites infecting the different primate hosts (Otto et al., 2018). Both a gene transfer event (Sundararaman et al., 2016) and an ancient introgression involving some erythrocyte invasion and surface protein genes (Otto et al., 2018) may have laid the groundwork for *P. falciparum's* host switch to humans. There is strong evidence of a horizontal transfer of an 8 kb region of chromosome 4 from *P. adleri* to the ancestor of *P. falciparum,* although disagreements exist as to whether the 8 kb fragment containing *Rh5* resulted from introgression or a horizontal transfer event (Sundararaman et al., 2016; Galaway et al., 2019). This region contained two erythrocyte invasion genes encoding the reticulocyte-binding-like homologous protein 5 (RH5) and the cysteine-rich protective antigen (CyRPA). RH5 differs from

the other members of the RH family in that it is a secreted rather than a membrane-tethered protein and cannot be genetically deleted in any *P. falciparum*. A combination of ancestral sequence reconstruction and quantitative protein interaction assays was used to investigate the role of RH5 in the emergence of *P. falciparum* in humans (Galaway et al., 2019). Probabilistic-based approaches to deduce the likely sequence of the *rh5* gene in the common ancestor of *P. falciparum* and *P. praefalicarum* converged on a single sequence, which the authors termed the ancestral introgressed *Rh5* gene (*IntRh5*). *P. falciparum* RH5 and its orthologs were expressed as soluble recombinant proteins, and the binding affinities of RH5 orthologs for human, gorilla, and chimpanzee basigins were quantified using surface plasmon resonance (SPR) and confirmed with cell-based binding assays. Surprisingly, IntRH5 protein was found to bind to human and gorilla basigin with similar affinities. This promiscuous receptor-binding phenotype was also observed in all parasites belonging to clade A (*P. adleri* and *P. gaboni*) but not in parasites belonging to clade B (*P. billcollinsi*, *P. reichenowi*, *P. preafalciparum*). As *P. falciparum* is restricted to humans with no African apes having been found to be naturally infected with this parasite, the question remains as to how this restriction occurred if the ancestral character state was permissive. Others have shown that recombinant PfRH5 does not interact with gorilla basigin, and only weakly interacted with chimpanzee basigin (Wanaguru et al., 2013). Galaway et al. (2019) identified six mutations separating *IntRh5* and *PfRh5*. A single mutation at residue 200 resulted in the complete loss of PfRH5 binding to gorilla basigin, suggesting that the host switch from gorillas to humans involved at least two steps, one in which erythrocytes from both hosts could be infected, followed by restriction in humans (Otto et al., 2018).

The erythrocyte binding-like (EBL) gene family also plays a critical role in erythrocyte invasion through the recognition of glycophorins, the predominant sialoglycoprotien on the erythrocyte surface. Different EBLs recognize different gylcophorins; for example, *Pf*EBA175 recognizes glycophorin A, and invasion of human erythrocytes by *P. falciparum* is heavily reliant on this EBL. The potential for sialic acid recognition to play a role in the exclusion of *P. falciparum* from chimpanzees and *P. reichenowi* from humans was recognized and investigated prior to the discovery of multiple *Laverania* species in African apes (Martin et al., 2005). The *P. falciparum* EBL165 gene (*PfEbl165*) is a pseudogene due to a frame-shift mutation in all known *P. falciparum* isolates studied so far (Proto et al., 2019). In contrast, all *Ebl165* homologs in ape *Laverania* species encode an intact protein (Otto et al., 2014b, 2018; Sundararaman et al., 2016). Due to a mutation in the gene encoding cytidine monophospho-N-acetylneuraminic acid hydroxylase (CMAH), human erythrocytes produce only Neu5Ac sialic acids whereas ape erythrocytes produce both Neu5Ac and Neu5Gc. EBA165 binding is dependent on the specific sialic acid repertoire displayed, and the absence of ape-specific Neu5Gc sialic acid on human RBCs is sufficient to prevent PfEBL165 from binding to them (Proto et al., 2019). CRISPR-Cas9 mediated correction of the frameshift mutation in *PfEbl165* led to the epigenetic silencing of a small number of genes in two regions of chromosome 4 and 11 that included *PfEbl165* itself and *PfRh4* (another invasion gene) (Proto et al., 2019). It appears as though a functional PfEBL165 is not compatible with robust growth of *P. falciparum* in humans, suggesting that the inactivation of this gene may have contributed to the parasites' establishment in this host. As an increasing number of candidate loci are studied, it is becoming clear that a single molecular event is not sufficient to explain the emergence of *P. falciparum* as a human pathogen. Accumulating evidence suggests that RH5, EBL165, other members of the RH5 complex, invasion proteins, and possibly exported proteins that remodel the infected erythrocyte surface all contribute to host specificity and tropism. Gene transfers between *Plasmodium* species have played an important role in the parasite evolution and host-parasite interactions (Plenderleith et al., 2019).

HOST-PARASITE INTERACTION AND EVOLUTION

Highly Polymorphic Immune Targets

Malaria parasites trigger an immune response the moment when they enter a host. To survive in this hostile environment, the parasite displays a range of strategies to evade host killing mechanisms, including variations in antigen epitopes targeted by host immune machinery and interference or suppression of specific arms of the host immune response (Recker et al., 2011; Yam and Preiser, 2017; Larsen et al., 2018; Xia et al., 2018). One of the consequences of these host-parasite interactions is increased genetic diversity at genes encoding proteins under host immune selection (**Figure 3A**), leading to genetic signatures of diversifying selection in the parasite genome. Some highly polymorphic genes include those encoding proteins such as the apical membrane antigen 1 (AMA1), merozoite surface protein 1 (MSP1), and circumsporozoite protein (CSP). These highly polymorphic genes as well as genome-wide polymorphisms such as single nucleotide polymorphisms (SNPs) and microsatellites (MSs) have been used for genotyping parasite strains, tracking parasite migration and disease outbreak, studying parasite molecular evolution, and evaluating host immune response and vaccine efficacy (Ranjit and Sharma, 1999; Volkman et al., 2001, 2007; Mu et al., 2002; Wootton et al., 2002; Joy et al., 2003; Jeffares et al., 2007; Ouattara et al., 2010; Drew et al., 2012; Tanabe et al., 2015; Dewasurendra et al., 2018; Stone et al., 2018). The high levels of genetic polymorphisms in immune targeted genes also provide signatures to identify potential vaccine candidates, particularly signatures of balancing selection within parasite populations (Mu et al., 2007; Tetteh et al., 2009; Ochola et al., 2010; Weedall and Conway, 2010; Amambua-Ngwa et al., 2012; Conway, 2015). For example, Tetteh et al. (2009) compared methods for prospectively identifying genes under balancing selection using 26 genes known or predicted to encode merozoite surface-exposed proteins and concluded that significant evidence of balancing selection could be detected using Hudson-Kreitman-Aguade (HKA) and Tajima's D (TjD) tests. In another study, high throughput sequencing of 65 clinical isolates from an endemic Gambian population identified 337

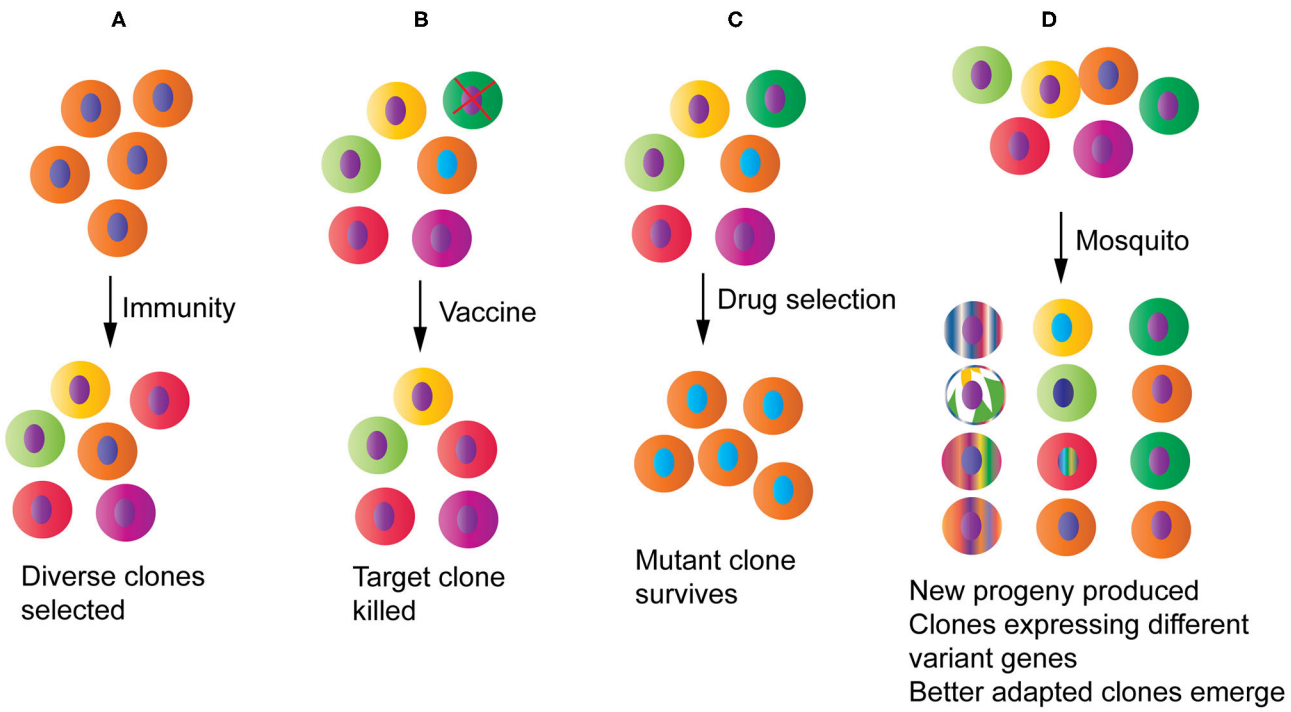


FIGURE 3 | Potential impacts of host-parasite interactions and human intervention measures on parasite populations and evolution. **(A)** Host immunity will pressure parasites to diversify and/or express different sets of variant antigen genes, increasing diversity at immune targets. **(B)** A partially protective vaccine may selectively remove a specific parasite strain from the population. **(C)** Drug treatment will reduce population diversity by removing the drug sensitive parasites and selecting for one or a few drug resistant parasite clones that may then spread to many endemic regions. **(D)** Parasites passing through mosquitoes may generate new progeny through genetic recombination and chromosomal reassortment, reset gene expression profile particularly those related to parasite survival, and select for clones that are better adapted to specific mosquito species.

genes with at least three SNPs that had Tajima's D values > 0, including 25 genes with Tajima's D values > 1.5 (Amambua-Ngwa et al., 2012). However, the use of highly polymorphic antigens as vaccines for protection against malaria is challenging because antibody responses to these molecules are generally allele-specific or strain-specific (Graves et al., 2016; Laurens et al., 2017), which represents one of the major roadblocks for developing an effective malaria vaccine.

Diversity and Evolution of Gene Families of Malaria Parasites

The *var* Genes

In addition to highly polymorphic antigen genes, there are many polymorphic gene families in the *Plasmodium* genomes such as the *P. falciparum var* genes, SICA (Schizont Infected Cell Agglutination) *var* (SICAvar) genes in *P. knowlesi*, and the *Plasmodium* interspersed repeat (*pir*) multigene family, which include *P. falciparum* repetitive interspersed (*rif*) and subtelomeric variant open reading frame (*stevor*), found in all *Plasmodium* species (Su et al., 1995; Cunningham et al., 2010; Jemmely et al., 2010; Wahlgren et al., 2017; Galinski et al., 2018). Indeed, large multigene families are present in many *Apicomplexa* parasites (Reid, 2015). These gene families evolve at high rates and play critical roles in antigenic variation and immune evasion. The *var* genes encode *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) proteins (Howard et al., 1988; Baruch et al., 1995; Su et al., 1995). Switches in expression among an estimated 60 *var* genes in every single parasite have been shown to correlate with variation in antigenic determinants that bind to different host receptors such as ICAM-1, CD36, EPCR, and other receptors (Smith et al., 1995; Wahlgren et al., 2017; Jensen et al., 2020). The binding of PfEMP1 proteins to host receptors is responsible for cytoadherence of infected erythrocytes (iRBCs) and the pathogenicity of severe malaria, particularly cerebral malaria (Smith et al., 2013; Jensen et al., 2020). The *P. falciparum* parasite evades host immunity via mutually exclusive expression (e.g., only one PfEMP1 is expressed on the surface of an iRBC at a time) of the highly diverse *var* family representing an almost unlimited gene pool at the parasite population level (Su et al., 1995; Chen et al., 1998; Hviid and Jensen, 2015; Deitsch and Dzikowski, 2017). Switches of *var* gene expression have been demonstrated in infections of human volunteers (Peters et al., 2002; Bachmann et al., 2019). Several mechanisms to explain the mutually exclusive *var* expression have been reported, including: (1) cis-acting DNA elements (promoters, silencers, activators) and anti-sense RNA transcripts acting to silence or activate the *var* genes; (2) epigenetic regulation through the presence or absence of

specific epigenetic marks; (3) locations of the *var* genes in specific subnuclear compartments; (4) coordination between members of the *var* family to ensure activation of a new gene while simultaneously silencing of the previously active gene at the same time (Deitsch and Dzikowski, 2017). Similarly, there are also several potential mechanisms for the generation of new *var* genes and maintenance of the *var* gene repertoire. Although each parasite carries a set of approximately 60 *var* genes, a parasite strain usually has a unique set of *var* genes with diverse sequences that differ from those of other strains (Trimnell et al., 2006; Chen et al., 2011; Claessens et al., 2014; Day et al., 2017; Ruybal-Pesantez et al., 2017). In fact, it was reported that every child infected with *P. falciparum* had a unique *var* DBLα repertoire (Ruybal-Pesantez et al., 2017), and some specific types of homologous blocks or motifs have been associated with severe diseases (Avril et al., 2013; Rorick et al., 2013). Considering frequent mixed infection of multiple parasite strains in an individual patient in endemic regions (Conway et al., 1991; O'Brien et al., 2016; Zhu et al., 2019), the *var* gene repertoire in the parasite populations of an individual patient or in a village can be quite large.

What are the forces or mechanisms driving the high diversity of the *var* genes? Since the PfEMP1 proteins are expressed on the iRBC surface, they are expected to be under strong immune selection. In addition to switches in expression to different *var* genes, one obvious mechanism to evade immune recognition is substitutions of critical amino acids at antigenic epitopes. A second mechanism for generating *var* gene diversity is the production of hybrid copies of *var* genes through genetic recombination during meiosis in the mosquito host. If genetic recombination is the major driving force, then sequencing parasite strains in mixed infections from an individual patient would reveal shared haplotypes and breakpoints in some *var* genes among parasite strains from a mixed infection. Interestingly, transmission through a mosquito also resets the *var* gene switching process (Peters et al., 2002), suggesting that going through mosquitoes or a sexual cycle can change *var* gene expression profile and may also alter *var* gene composition in the genome at the same time. A third mechanism is generation of recombinant *var* genes through somatic recombination during asexual replication. Recombination events such as duplicative transposition of specific sequences were observed during parasite transformation in *in vitro* culture (Frank et al., 2008). By constructing large parasite clone trees and performing whole genome sequence analysis of *var* gene sequences in asexually replicating parasites, Claessens et al. (2014) found that the *Var* exon 1 recombined in a rate up to 0.2% of infected erythrocytes *in vitro* per life cycle, suggesting that *var* gene sequence polymorphism could be mainly generated during the asexual part of the life cycle (Claessens et al., 2014). Using targeted DNA double-strand breaks (DSBs) and long-read whole-genome sequencing, Zhang et al. (2019) showed that a cascade of recombination events could occur starting at a single DNA break within a subtelomeric region generating multiple new *var* genes (Zhang et al., 2019).

What is the mechanism mediating somatic recombination? One possible way of generating recombinant *var* genes during the asexual cycle is through microhomology-mediated end joining (MMEJ), involving alignment of microhomologous sequences flanking a DNA break and subsequent removal of the intermediate sequences (Grajcarek et al., 2019). Because of the absence of enzymes necessary for non-homologous end joining in *Plasmodium*, repair of DNA breaks generally required homologous templates (Lee et al., 2014). However, an end joining pathway for repair of DSBs without the presence of homologous sequence was observed in *P. falciparum*, with the repair junctions frequently containing short insertions in the surrounding sequences (Kirkman et al., 2014), suggesting that MMEJ may play a role in these events. Various enzymes for MMEJ are indeed present in *P. falciparum* parasites (Lee et al., 2014). Recently, MMEJ was shown to function in the *P. yoelli* genome and was able to generate a large number of deletion or insertion mutants of varying sizes in the central repeat region of *P. yoelii* CSP gene (Xu et al., 2019). Therefore, homologous regions among the different copies of the *var* genes may act as templates during DSB repair, leading to deletion, insertion, and translocation of repetitive DNA sequences and rapid evolution of new *var* genes.

The *var* genes have a GC content much higher than those of other protein coding genes in the *P. falciparum* genome (∼34% vs. 20%), and how the genes with high GC content originated and are maintained in a genome with a high AT content remains largely unknown (Su et al., 1995; Gardner et al., 2002). One potential explanation is that the *var* gene came from the parasite hosts (humans and mosquitoes) through horizontal gene transfer; however, there is no evidence that the host genomes contain genes with structures and/or protein motifs observed in the PfEMP1. The *var* genes are not even present in the other human malaria parasites such as *Plasmodium vivax,* whose genome has high GC content in its genome (Carlton et al., 2008; Chan et al., 2015). Because *var* genes contain Duffy binding-like (DBL) domains that are also present in another gene family encoding erythrocyte invasion ligands (EBL proteins), it has been proposed that the *var* genes might have originated from the EBL genes (Smith et al., 2013). The *var* genes are also present in the *Laverania* parasites infecting African great apes (*P. praefalciparum, P. blacklocki, P. adleri, P. billcollinsi, P. gaboni,* and *P. reichenowi*) (Rask et al., 2010; Otto et al., 2014b, 2018; Laurens et al., 2017). Therefore, the *var* genes were already present when *P. falciparum* diverged from *P. preafalciaprum* to emerge as a human parasite 40,000–60,000 years ago (Otto et al., 2018). The question as how the *var* genes initiated in the ancestor remains unanswered. If *var* genes were derived from the genes containing DBL domains, why similar gene family expansion events did not occur in other malaria parasite species?

The *pir* Genes

Similar to the *var* genes, the *pir* multigene family is likely under host immune pressure and plays an important role in immune evasion (Cunningham et al., 2010). The *pir* gene family derives its name from the first initial of the parasite species name followed by "ir," an abbreviation for interspersed repeat: *yir* for *P. yoelii* genes; *bir* for *P. berghei* genes; *vir* for *P. vivax* genes, and so on

(Cunningham et al., 2010). Gene families of *stevor*, *rif*, and PfMC-2TM in *P. falciparum* were classified as *pir* gene families because of similar gene structures and sequence similarity in some introns (Janssen et al., 2002). The *pir* genes have a conserved three-exon structure, including a short first exon, a long second exon encoding conserved cysteine residues and a transmembrane domain, and a conserved third exon encoding cytoplasmic domain (Janssen et al., 2002). The *pir* genes are mostly distributed in the subtelomeric regions of chromosomes (except *kir* in *P. knowlesi*) with gene copies numbering from a few dozen to more than 800 (Cunningham et al., 2010). Distinct from the *var* genes, up to 40% of the *cir* gene repertoire are expressed throughout the intraerythrocytic cycle of development of the *P. chaboudi* parasite during infection (Lawton et al., 2012). The timing and level of transcription differs between *cir* genes, with ring, trophozoites, and schizonts expressing different sets of major *cir* genes (Cunningham et al., 2009). Additionally, individual CIR proteins were observed within differential localizations of iRBCs, including on the surface of iRBCs and merozoites, some of which could bind mouse erythrocytes and play a role in immune evasion (Yam et al., 2016; Yam and Preiser, 2017).

The mechanisms for generating *pir* gene diversity are expected to be similar to those of *var* genes, including amino acid substitutions at specific antigen epitopes in response to immune pressure and the generation of new recombinant genes through genetic recombination as well as MMEJ mediated somatic deletion, insertion and translocation. Genes with mosaic sequences or two genes with segments of 100% identity in the *P. yoelii* genome have been noted, suggesting crossovers between two copies of *pir* genes (Cunningham et al., 2010). However, rodent malaria parasites may have a lower degree of antigen diversity and undergo significantly less ectopic recombination than *P. falciparum,* partly due to the lack of components of the translesion (TLS) polymerases that are required for the recombination events that drive diversification of the multicopy gene families (Kirkman and Deitsch, 2020; Siao et al., 2020). Further discussions on functions of the *pir* genes can be found here (Cunningham et al., 2010; Su et al., 2019). More studies are necessary to better understand the mechanisms of gene expression regulation and the maintenance of genetic diversity of the *pir* genes.

The EBL Genes and Evolution of Virulence in Laboratories

One of the consequences of host immune pressure on parasites is the generation of parasites with altered virulence. Whereas it has been difficult to evaluate how host immune response or vaccination impacts parasite evolution and virulence in humans, changes in parasite growth and virulence can be measured in rodent malaria parasites in the laboratory. One example is the *P. yoelii* EBL protein (PyEBL) that is known to bind the Duffy antigen/receptor for chemokines (DARC), a transmembrane glycoprotein expressed on epithelial cells, endothelial cells, and erythrocytes (Swardson-Olver et al., 2002; Woolley et al., 2005; Culleton and Kaneko, 2010). Whereas several genes encoding proteins with DBL domains have been found in *P. falciparum* such as PfEBL-140, PfEBL-175, and PfEBL-181, there is only a single copy of the EBL gene in *P. yoelii* (Culleton and Kaneko, 2010). PyEBL contains a single DBL domain, unlike other EBLs which may have two DBL domains per molecule.

Immunity or host response to malaria infection is generally parasite strain specific (Fluck et al., 2004; Cheesman et al., 2006; Early et al., 2018). The principle of strain specific immunity has been applied to identify immune targets in rodent malaria parasites using a strategy called linkage group selection (LGS) (Culleton et al., 2005, 2011; Pattaradilokrat et al., 2009; Cheesman et al., 2010; Abkallo et al., 2017). In a LGS analysis, phenotype-specific selection pressure such as immunity is applied to uncloned progeny of a genetic cross between two malaria parasites that differ in phenotype (strain specific immunity in this case) (Culleton et al., 2005; Cheesman et al., 2010). PyEBL has also been identified by LGS experiments because it plays a role in parasite invasion of RBCs and is also the target of host immunity (Pattaradilokrat et al., 2009; Abkallo et al., 2017). Mutations in the PyEBL gene not only affect parasite growth and virulence, but also influence iRBC surface protein and host immune response (Otsuki et al., 2009; Peng et al., 2020). There are isogenic pairs of rodent malaria parasites that have different parasite growth rates in mice and produce very different diseases. *P. yoelii yoelii* 17X (or YM) infection of BALB/c or C57BL/6 mice causes host death within 7 days post infection (pi), whereas its isogenic strain *P. y. yoelii* 17XNL (or 17X) is cleared by the host around 21 days pi (Otsuki et al., 2009; Pattaradilokrat et al., 2009). Both parasites were derived from *P. yoelii* 17X during passages in separate laboratories (Pattaradilokrat et al., 2008). A replacement of Cys to Arg (at position 731 or C731R) at the second Cys position in region 6 of PyEBL was found to be the major determinant for the difference in virulence and parasite growth; however, this mutation cannot explain the entire parasite growth phenotype (Otsuki et al., 2009). A second example is the isogenic parasites of *P. y. nigeriensis* N67 (N67) and *P. y. nigeriensis* N67C (N67C) that also have very different disease phenotypes in C57BL/6 mice, including parasitemia, tissue pathology, and host mortality (Wu et al., 2014; Lacerda-Queiroz et al., 2017). Parasitemia in C57BL/6 infected with N67 (1×10^6 iRBCs) increases to ∼40–50% day 5 pi, declines to below 10% on day 7 pi, and increases to ∼60% at day 15 pi (Wu et al., 2014). The N67 infected mice die at ∼day 20 pi. Mice infected with N67C have ∼50% parasitemia day 6 pi and begin to die day 7 pi with declining parasitemia (Lacerda-Queiroz et al., 2017). Similarly, a substitution of Cys to Tyr (C741Y) in the protein trafficking domain of PyEBL between N67 and N67C parasites was found to influence disease phenotypes and host mortality (Peng et al., 2020). The substitution alters processing and trafficking of the PyEBL protein, reduces PyEBL binding to Band 3 of the RBC, increases phosphatidylserine (PS) surface exposure on iRBCs, and enhances iRBC osmotic fragility. Interestingly, the two studies have different conclusions on the mechanism of how the PyEBL substitutions affecting disease phenotypes, although the C731R and C741Y substitutions both appear to change PyEBL protein trafficking, from microneme localization to dense granules in the cytoplasm. For C731R substitution between 17X and 17XL (YM), parasite virulence was mostly attributed to altered erythrocyte invasion (Otsuki et al.,

2009). In contrast, in the C741Y substitution in PyEBL between N67 and N67C was found to modulate host recognition of iRBCs and immune responses through interaction with RBC membrane molecules such as Band 3 and PS (Peng et al., 2020). A third example is the isogenic parasites *P. berghei* ANKA and *P. berghei* NK65: *P. berghei* ANKA infection is lethal with cerebral malaria symptoms and has been used as an experimental cerebral malaria (ECM) model, whereas *P. berghei* NK65 infection generally does not have neurological symptoms (Lacerda-Queiroz et al., 2011; Nacer et al., 2014). Even cloned lines of *P. berghei* ANKA can differ in their ability to induce ECM in the same host (Amani et al., 1998). However, the molecular mechanisms responsible for the differences in disease phenotype and virulence among these parasites are not clear. Additionally, the substitutions in PyEBL are likely not the only genetic changes contributing to the total disease phenotypes. Unknown differences including SNPs and variation in genes such as the *pir* multigene gene family may also contribute the differences in disease phenotypes and virulence (Cunningham et al., 2010). These parasite pairs are examples of the evolution of parasite genomes and changes in virulence following laboratory passages. There are additional parasite species and strains, including *P. chabaudi* and *P. vinckei*, that showed changes in parasite genome and virulence (Perkins et al., 2007). It can be expected that these types of phenotypic changes in parasite growth, disease virulence, and drug response are occurring daily in the huge parasite populations in endemic regions of the world. Additionally, mixed infections of multiple *P. falciparum* strains may lead to higher proportion of severe anemia, pulmonary complications, and multiple organ failure than those with single-strain *P. falciparum* infection (Kotepui et al., 2020). The dynamics of parasite genetic variation and gene expression are major roadblocks for developing effective vaccines and therapies.

PARASITE SELECTION ON THE HOST GENOME

Whereas host immunity is one of the major factors driving parasite evolution, malaria parasites and many other pathogens also have tremendous impacts on the adaption and evolution of their hosts (Daub et al., 2013). The well-known Haldane malaria hypothesis suggests an evolutionary advantage of the thalassemic condition in protection against malaria (Akide-Ndunge et al., 2003; Canali, 2008). Many proteins expressed on the RBC surface such as sickle cell HbS variants, DARC, and glycophorin A (GYPA), and proteins that play a role in host immune response such as TNF-α and IFN-γ have been associated with protection from severe complications of malaria (Carter and Mendis, 2002; Kwiatkowski, 2005; Verra et al., 2009; Hedrick, 2012; Kariuki and Williams, 2020). The protective genes were mostly identified through association studies including genome-wide association studies (GWAS). For example, a recent GWAS study performed in a Tanzanian population revealed association of protective roles of sickle cell HbS variant, interleukin receptors (IL-23R and IL-12RBR2), and the kelch-like protein KLHL3 (Ravenhall et al., 2018). In another study, Ebel et al. analyzed mammalian adaptation in 490 *Plasmodium*- or Piroplasm-interacting proteins (PPIPs) and showed that blood parasites imposed specific selection on PPIPs, leading to a higher number of adaptive substitutions in the PPIPs than expected throughout mammalian evolution (Ebel et al., 2017). Interestingly, many immune PPIPs were also enriched for viral and bacterial interactions, with 48% of the PPIPs interacting with viruses and/or bacteria. Moreover, there was a 2.5X enrichment of adaptation when only PPIPs interacting with *Plasmodium* or Piroplasms were considered. Malaria parasites interact with a large number of host genes (Wu et al., 2015, 2020), which may enhance adaptive substitutions in these genes. There are many excellent reviews on the effects of parasite infections, particularly malaria parasite infections, on host genome evolution (Tishkoff and Williams, 2002; Kwiatkowski, 2005; Mangano and Modiano, 2014; Goheen et al., 2017).

IMPACT OF VACCINES, DRUGS, AND VECTORS ON PARASITE POPULATIONS AND DISEASE

Types of Malaria Vaccines

Developing an effective vaccine has been one of the major goals of malaria research for more than 40 years, and we are still far away from achieving the goal. Malaria vaccines can be divided into several categories based on the parasite developmental stages such as pre-erythrocytic vaccines, blood stage vaccines including those against pregnancy-associated malaria, transmission-blocking vaccines, multistage or multiantigen vaccines, and whole organism based vaccines (Frimpong et al., 2018; Laurens, 2018; Beeson et al., 2019). Development of a cross-species vaccine that protects against multiple *Plasmodium* species or strains is an important goal. *P. falciparum* 3D7 strain has been the model strain for studies of genome structure and analyses of transcriptome and proteome. Subsequently, the majority of vaccines are based on sequences from the 3D7 parasite; however, the 3D7 strain is often not the most prevalent strain in malaria-endemic areas. Currently, the leading vaccine candidates are mostly based on polymorphic antigens with the vaccine target representing only a single variant strain. However, immunity to malaria is generally strain-specific, and the development of an effective vaccine is often hampered by strain-specific and/or partial immunity generated by vaccination.

Partially Protective Vaccines and Effects on Host Populations

The most promising and extensively tested malaria vaccine candidate against *P. falciparum* parasites is RTS,S/AS01 that consists of 18 copies of the central repeat and the C-terminal domain of PfCSP fused to hepatitis B virus surface antigen (HBsAg) (Draper et al., 2018). In a large-scale phase III trial involving 15,460 children, including 6,537 infants aged 6–12 weeks and 8,923 children aged 5–17 months, vaccine efficacy (VE) against clinical malaria was reported to be ∼50–56%, and VE against severe malaria in the combined age children was ∼35% in the per-protocol population at 11 month follow-up

(Rts et al., 2011). However, in a 7 year follow-up analysis, the VEs of RTS,S/AS01 were estimated to be 35.9% in the first year after vaccination, 2.5% in the fourth year, and a negative VE in the fifth year (Olotu et al., 2016). More generalized convulsive seizures were observed in the RTS,S vaccinated group than the control group (comparator rabies vaccine group) in the older children (5–17 months) within 7 days after RTS,S/AS01 vaccination (Rts et al., 2011). Additionally, significantly more cases of meningitis were reported in the older children (Rts et al., 2012; Rts, 2015). Moreover, RTS,S vaccine was later found to be associated with significantly higher all-cause mortality in girls of both age groups (Klein et al., 2016). The higher rates of meningitis and mortality in girls have prompted the World Health Organization (WHO) to issue a position paper cautioning on malaria vaccines, with emphasis on the RTS,S/AS01 (World Health Organization, 2018) and discussions on the need of further studies on vaccine safety and negative non-specific effects (Aaby et al., 2015; Muller et al., 2015). Higher all-cause mortality was also observed in girls after vaccination with the high-titer measle vaccine (HTMV) (Aaby et al., 1993), which was later found to be likely associated with diphtheria-tetanus-pertussis (DTP) vaccine or interaction between HTMV and DTP (Aaby et al., 2003, 2007). Generally, women have a stronger cellular and humoral responses to infections and are more likely to have autoimmune diseases due to X-linked genes, hormones, and societal context (Fish, 2008; Billi et al., 2019). The observation of higher rate of severe symptoms in RTS,S vaccinated girls could be due to enhanced immune reactions from vaccine priming of the immune system.

A partial protective vaccine with enhanced adverse effects brings out some important issues in malaria vaccine development that has largely been ignored so far: Can a partially protective malaria vaccine prime for more severe disease symptoms and select for parasite population "resistant" to specific vaccines? The T-cell prime hypothesis proposed to explain clinical malaria and antibody dependent enhancement (ADE) observed in many viral infections may provide some answers to the questions (Porterfield, 1986; Riley, 1999). More than 20 years ago, Dr. Eleanor Riley proposed that T-cell priming is required for amplification of the inflammatory response to malaria, which can explain the patterns of clinical malaria in both endemic and non-endemic populations (Riley, 1999). According to the hypothesis, primary infections would induce low levels of IFN-γ and TNF-α; at the same time, antigen-specific T cells are primed by the infection. On reinfection, large amounts of cytokines such as IFN-γ and TNF-α are produced by the primed T cells, leading to an increased risk of cerebral malaria or other severe symptoms. She also cautioned that vaccinations with BCG or tetanus toxoid might prime T cells to respond to malaria antigens and suggested that having an appropriate balance between protective and pathogenic levels of Th1-derived cytokines will be crucial for a successful outcome of a vaccine (Riley, 1999). A partial protective vaccine may also prime the immune system in some individuals that can mount a strong inflammatory response when infected, leading to enhanced severe symptoms. Another related possibility is ADE that was observed in various viral infections, including Dengue virus (Waggoner et al., 2020), influenza (Winarski et al., 2019), Zika virus (Asad et al., 2019), coronavirus (Negro, 2020; Wan et al., 2020), and *Acinetobacter baumannii* bacteria (Wang-Lin et al., 2019). ADE is a well-known mechanism that viruses or bacteria may infect susceptible cells by interacting with antibodies or complement components (Fc or complement receptors), leading to enhanced viral entry into host cells and/or higher replication rate. Cross-reactive IgGs can opsonize viral particles by binding the virus through Fc receptors on host cell surface and may also mediate immune suppression to further increase viral load (Halstead et al., 2010). Monoclonal antibodies have been shown to destabilize the influenza HA stem domain, leading to faster kinetics of influenza virus fusion and enhanced respiratory disease (Winarski et al., 2019). Antibody enhancement of malaria transmission and complement activation in conjunction with antibodies in enhancing RBC invasion have been reported too (Biryukov et al., 2016; Stone et al., 2019). The RTS,S vaccine might prime T cells in some hosts and/or produce antibodies that can amplify immune response and enhance clinical symptoms. Addition of an arm to evaluate the potential of a vaccine to enhance adverse effects should be included before or during a vaccine trial.

Effects of Vaccination on Parasite Population and Gene Expression

Malaria parasite populations in endemic regions consist of a large number of strains that express different alleles of vaccine target proteins. Due to technical limitation, malaria vaccines are usually designed based on sequences from one to a limited number of alleles of a vaccine target. It is has been shown that anti-malaria immunity is mostly strain-specific (Cheesman et al., 2006; Graves et al., 2016; Laurens et al., 2017; Early et al., 2018), and a vaccine based on a limited number of alleles of a target may select for parasite populations with alternative alleles. Indeed, vaccination with pneumococcal conjugate vaccines (PCVs) changed the pneumococcal populations in children, resulting in statistically significant shift from vaccine-type population to non-vaccine-type populations (Quirk et al., 2018). Similarly, using the rodent malaria model *P. chabaudi* and recombinant AMA-1 antigen, it was shown that mono-allelic immunization increased the frequency of heterologous clones in mixed clone infections (Barclay et al., 2008). Moreover, analysis of parasite genotypes collected from the RTS,S phase III trial showed that the RTS,S vaccine had greater activity against malaria parasites with the matched PfCSP allele than against those with mismatched alleles (Neafsey et al., 2015). Therefore, vaccination with an allele of the target antigen will likely change target allele proportion in parasite populations, and large-scale vaccination may lead to vaccine mediated depletion of specific alleles targeted by the vaccine (**Figure 3B**). The impacts of this type of selection on parasite populations, including the possibility of selecting more virulent parasite strains, remain largely unknown and required further investigations. Interestingly, a vaccine may also alter parasite expression of variant antigen genes. Vaccination of 10 African volunteers with the PfSPZ vaccine showed parasites from individuals with intermediate antibody levels expressed only few *var* gene variants, whereas those with

low antibody levels expressed a broad spectrum of multiple, predominantly subtelomeric *var* genes coding PfEMP1 binding to endothelial protein C receptor (EPCR) that is associated with severe childhood malaria (Bachmann et al., 2019). Therefore, vaccination certainly can impact parasite population dynamics, genetic diversity, and possibly virulence.

Drug Selection on Parasite Populations and Evolution

In addition to host immune response, antimalarial drugs have played a significant role in shaping parasite populations. Application of antimalarial drugs puts pressure on parasite populations and likely will select parasites with mutations that can confer resistance to the drugs, leading drug selective sweeps (**Figure 3C**). Some examples of drug selective sweep include mutations in *P. falciparum* chloroquine resistance transporter (PfCRT) and in dihydrofolate reductase-thymidylate synthase and dihydropteroate synthase (PfDHFR-TS and PfDHPS) conferring resistances to chloroquine and pyrimethamine/sulfadoxine (PS), respectively (Peterson et al., 1990; Fidock et al., 2000). Under drug pressure, a small number of parasites with resistant mutations will survive, whereas parasites without the mutations are killed by the drugs. With large-scale drug use, parasites with resistant mutations will spread, replacing parasites sensitive to the drugs. Indeed, worldwide chloroquine and PS selective sweeps have been reported (Wootton et al., 2002; Roper et al., 2004; McCollum et al., 2008), which can greatly reduce population diversity in many endemic areas generating population bottlenecks. Drug selective sweeps are important factors in shaping the current worldwide parasite populations and may have contributed to Malaria's Eve hypothesis debate (Rich et al., 1998; Su et al., 2003). The sweeps of parasites derived from one or a few mutants conferring resistance to a drug will result in a relatively homogeneous parasite population (**Figure 3C**), as was seen in the Brazilian Amazon after wide spread use of chloroquine (Wootton et al., 2002). A relatively homogeneous parasite population may be advantageous for vaccine development.

The Role of Mosquito Vectors in Altering Parasite Populations

Mosquito vectors can shape parasite populations in many ways (**Figure 3D**). First, genetic changes such as chromosome reassortment and crossover will occur inside a mosquito and generate new parasite strains if a patient is infected with two or more variant strains, which will increase parasite population diversity. A high multiplicity of infection is frequently observed in patients in regions with high transmission intensity (Kobbe et al., 2006; Felger et al., 2012). Second, parasite transmission through mosquitoes can also re-program gene expression profiles. For example, changes in composition and frequency of *var* gene transcripts were observed between cultured *P. falciparum* parasites used to infect mosquitoes and the parasites recovered from infected volunteers after mosquito bites, suggesting re-programing *var* gene expression profile (Peters et al., 2002). Similarly, attenuated parasite virulence was found to be associated with modified expression of the *pir* multi-gene family. Transmission of *P. c. chabaudi* through the mosquito changes gene expression of the *pir* multi-gene family in the erythrocytic cycle, leading to changes in parasite virulence and immune response in the mammalian hosts (Spence et al., 2013). Third, vector competency can also influence parasite transmission frequency. A study of *P. vivax* populations in southern Mexico showed that the distributions of parasite populations largely mirrored those of mosquito species, which was confirmed by feeding experiments in the laboratory (Joy et al., 2008). The results suggest a reciprocal selection between malaria parasites and mosquito vectors and local adaptation of the parasites to mosquito species in different environments (coastal or hill sites). *P. falciparum* isolates from Africa, Asia, or the Americas were shown to have low compatibility to malaria vectors from different continents, which is mainly mediated by a parasite protein called Pfs47 due to its interaction with mosquito immune system (Molina-Cruz et al., 2015). Mosquito phagocytes play an important role in mosquito vector competence and immunity against malaria infection (Kwon and Smith, 2019). Additionally, *Serratia marcescens* bacteria isolated from either laboratory-reared mosquitoes or wild populations in Burkina Faso shows great phenotypic variation that is directly correlated with its ability to inhibit *Plasmodium* development within the mosquito midgut (Bando et al., 2013).

APPROACHES FOR STUDYING HOST-PARASITE INTERACTIONS

Identification of Host-Parasite Interacting Genes Using Genetic Crosses

In addition to antibody response to antigens expressed on the surfaces of iRBCs, malaria infections also stimulate differential activation of immune cells and the production of cytokines and chemokines as part of the innate immune response. What are the parasite ligands that initially trigger host innate responses? What are the host receptors that recognize the parasite molecules? How are the host responses to malaria infections regulated? What are the impacts of the host-parasite interaction on the evolution of both the host and parasite genomes? As discussed above, host immunity drives diversity of parasite antigen genes, and parasite infections select for protective alleles in host response genes. Large numbers of parasite antigens and host protective genes have been identified (Kwiatkowski, 2005; Salinas et al., 2019; Vijayan and Chitnis, 2019; de Jong et al., 2020). However, the numbers of host and parasite molecules involved in host-parasite interactions are likely much larger than the numbers of genes so far investigated. For example, Wu et al. (2015) performed an immunogenetic screen using trans-species expression quantitative trait locus (ts-eQTL) analysis and significantly (LOD score $\geq$ 3.0) linked 1,054 host genes to a large number of parasite genetic markers/loci (Wu et al., 2015). In this study, microsatellite genotypes of the *P. yoelii* malaria parasite and phenotypes of host (mouse) gene expression levels day 4 pi were analyzed for significant linkages between parasite genetic markers and host gene expression levels. Additionally,

host genes that were linked (LOD score $\geq$ 3.0 or 2.0) to parasite genetic loci were clustered based on genome-wide patterns of LOD scores (GPLSs), showing that genes functioning in related pathways generally have similar GPLSs. This study demonstrates that a large number of host genes respond to *P. yoelii* parasites as early as day 4 pi. Similarly, many parasite genetic loci were predicted to interact with host genes belonging to specific pathways. For example, a genetic locus on parasite chromosome 8 was significantly linked to a large number of host genes play a role in hematopoiesis, suggesting that a gene or genes in the locus may influence host hematopoiesis. The study represents the first QTL analysis involving two organisms, microsatellite genotypes from malaria parasites and gene expression as phenotypes from the host, with the advantage of identifying interacting genes from both the parasite and the host. This ts-QTL analysis is a novel approach for identifying interacting genes from both parasite and the host at the same time, taking advantage of genetic recombination and chromosome reassortment during the sexual stage of a pathogen in generating progeny with unique genomes inherited from the parents. In a follow-up study using mRNA isolated from the spleens of infected mice 24 h pi (earlier than day 4), many host receptors including Toll-like receptors (TLR), retinoic acid-inducible gene-I-like receptors (RLRs), G-protein coupled receptors (GPCRs), and olfactory receptors (ORs) were significantly linked to parasite genetic loci, providing important candidate genes for studying host molecules that recognize parasite ligands or pathogen-associated molecular patterns (PAMPs) (Wu et al., 2020). Ts-QTL holds the potential to identify more interacting genes using additional genetic crosses or host RNA samples collected at different time points throughout the parasite life cycle.

Transcriptome Analysis of Patient Samples

Another approach to identify host gene responses to malaria infections is to use samples collected from malaria patients, although a study using patient blood samples can be influenced by many factors such as variations in infection time, host genetic background and immune status, mixed parasite infections, nutrition status, antimalarial drug usages, etc. Mixed strain infections can change parasite growth dynamics, relapse time (for *P. vivax*), disease severity, and host mortality rate (Lover and Coker, 2015; Kotepui et al., 2020). Similarly, the genetic composition of each patient is unique, and variations in host genetic background will also affect host responses. Therefore, dissecting host-parasite genetic interactions using patient samples has proven to be much more complicated than using infections in inbred mice with cloned progeny from a genetic cross.

As discussed above, GWAS has been successful in identifying some host genes that contribute to protection against malaria. As an alternative, analysis of host gene expression in response to malaria infection can be useful too, particularly when transcripts from both the host and parasites are analyzed at the same time using RNAseq (Greenwood et al., 2016). Using blood samples of *P. falciparum*-infected West African children, Idaghdour et al. (2012) performed a joint genome-wide analysis of gene expression and genetic polymorphisms, and showed that malaria infection and host genotype both influenced immune gene expression throughout the genome (Idaghdour et al., 2012). The study implicated many genes in the complement system, antigen processing and presentation, and T-cell activation in response to malaria infections (Idaghdour et al., 2012). For example, the authors noted that the expression of five genetically regulated human leukocyte antigen (HLA) class II loci was negatively correlated with parasite load and molecules such as IL18R1, TLR4, TLR5, IFNGR1, and IFNGR2, suggesting impairment of antigen processing and immune signaling. Impairment of antigen processing and T cell activation, including suppression of various MHCII genes, were also observed in C57BL/6 mice infected with different *P. yoelii* strains and progeny of a *P. yoelii* genetic cross (Xia et al., 2018). In a dual transcriptome analyses of the host and parasite genes in samples from 46 malaria-infected Gambian children, Lee et al. (2018) performed multivariate analyses and showed association of disease severity with increased expression of granulopoiesis and IFN-γ related genes as well as inadequate suppression of IFN-I signaling (Lee et al., 2018). In another analysis by the same group, lower expression of *cxcl10*, HLA genes, and IFN-I associated genes, and higher expression of cathepsin G and matrix metallopeptidase 9 genes were associated with parasite growth inhibition (Georgiadou et al., 2019). Moreover, a blood transcriptional analysis in Senegalese patients with cerebral malaria (CM), severe non-cerebral malaria (NCM), or mild malaria (MM) identified host gene clusters that discriminated between CM and MM patients, showing that genes involved in BCR-, TCR-, TLR-, cytokine-, FcεRI-, and FCGR- signaling pathways played a role in cerebral malaria (Thiam et al., 2019). Associations between disease severity and PfEMP1 transcripts and specific domains were also observed in RNAseq data from 44 parasite isolates that caused SM and MM malaria in Papuan patients (Tonkin-Hill et al., 2018). These studies illustrate that there are signatures of host transcriptional responses to malaria infections and disease severity. Indeed, blood transcriptome analyses of pathogen-specific host immune responses have been proposed as a tool for improving the diagnosis and classification of patients with infectious diseases (Mejias et al., 2014). However, among the studies of host transcriptional response to malaria infections in animal models and in humans, the results are not always consistent, likely due to differences in study design, study objective, transmission intensity, sample collection time, as well as lack of replication. Caution is warranted when interpreting the results from field-based analyses.

There are other approaches for studying host-parasite interaction such as analysis of genome diversity and genetic association (Early et al., 2018; Damena et al., 2019; Malaria Genomic Epidemiology Network, 2019), computational prediction of host-parasite interactomes (Wuchty, 2011; Cuesta-Astroz et al., 2019), metabolite profiling to investigate transcriptomic changes during malaria infection (Tang et al., 2018), and systems biology/integrative approaches to study host immune responses (Mahanta et al., 2018). These are just few emerging and promising examples using -omics data to study host-parasite interactions.

SOME KEY QUESTIONS AND CHALLENGES

There are several area that require special attention: (1) The new *Laverania* species are identified through nucleic acid sequence analyses. Morphological and biological characterization of the parasites will be critical for establishing these parasites as solid independent species. (2) The molecular mechanism of host specificity of the *Laverania* species is interesting and requires further study. (3) The origin of the *P. falciparum var* genes remains unknown. How did this gene family with a high GC-content originated and how is it maintained in a AT-rich genome? (4) Similar to the *var* gene family, the origins and functions of the *pir* gene families require additional studies. The *pir* genes likely play important roles in virulence and immune evasion. (5) More studies are necessary to elucidate the mechanisms that generate and maintain diversity of the *var* and *pir* genes in a parasite cell. (6) The impacts of antimalarial drugs and vaccines on the diversity and evolution of malaria parasite populations are important issues related to disease control and management. (7) The long-term impacts of antimalarial drugs and vaccines on host health and immune system also require more attentions. ADE is a real possibility for malaria vaccines. (8) Insect vectors play important roles in the evolution, diversity and transmission of malaria parasites. This is an area that has not received sufficient attention. (9) The parasite molecules that trigger host innate immune responses remain largely uncharacterized. DNA and RNA have been implicated in stimulating type I interferon responses, and strain-specific responses have been observed. However, the specific sequences that can bind to host receptors and the mechanisms of the strain-specific innate immune responses require additional studies.

CONCLUSION

Host-parasite interactions are the main forces driving the evolution of both malaria parasites and their hosts. Genome diversity and gene function have co-evolved shaping the current genomes of malaria parasites and their hosts. Application of antimalarial drugs, deployment of vaccines, modulation of host immunity, and the development in mosquito vector control measures all have the potential to alter parasite populations including genome diversity and virulence. Proper monitoring of parasite populations in the context of drug, vaccine, and vector control programs is necessary to minimize unintended negative impacts on parasite population dynamics and disease severity.

AUTHOR CONTRIBUTIONS

X-zS and DAJ conceive ideas and writing. CZ wrote and created **Figure 1**. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), USA.

ACKNOWLEDGMENTS

The authors thank Bradley Otterson, NIH Library Writing Center, for manuscript editing assistance.

REFERENCES

Aaby, P., Garly, M. L., Nielsen, J., Ravn, H., Martins, C., Bale, C., et al. (2007). Increased female-male mortality ratio associated with inactivated polio and diphtheria-tetanus-pertussis vaccines: observations from vaccination trials in Guinea-Bissau. *Pediatr. Infect. Dis. J.* 26, 247–252. doi: 10.1097/01.inf.0000256735.05098.01

Aaby, P., Jensen, H., Samb, B., Cisse, B., Sodemann, M., Jakobsen, M., et al. (2003). Differences in female-male mortality after high-titre measles vaccine and association with subsequent vaccination with diphtheria-tetanus-pertussis and inactivated poliovirus: reanalysis of West African studies. *Lancet* 361, 2183–2188. doi: 10.1016/S0140-6736(03)13771-3

Aaby, P., Knudsen, K., Whittle, H., Lisse, I. M., Thaarup, J., Poulsen, A., et al. (1993). Long-term survival after Edmonston-Zagreb measles vaccination in Guinea-Bissau: increased female mortality rate. *J. Pediatr.* 122, 904–908. doi: 10.1016/S0022-3476(09)90015-4

Aaby, P., Rodrigues, A., Kofoed, P. E., and Benn, C. S. (2015). RTS,S/AS01 malaria vaccine and child mortality. *Lancet* 386, 1735–1736. doi: 10.1016/S0140-6736(15)00693-5

Abkallo, H. M., Martinelli, A., Inoue, M., Ramaprasad, A., Xangsayarath, P., Gitaka, J., et al. (2017). Rapid identification of genes controlling virulence and immunity in malaria parasites. *PLoS Pathog.* 13:e1006447. doi: 10.1371/journal.ppat.1006447

Adler, S. (1923). Malaria in chinpanzees in Sierra Leone. *Ann. Trop. Med. Parasitol.* 17, 13–19. doi: 10.1080/00034983.1923.11684345

Akide-Ndunge, O. B., Ayi, K., and Arese, P. (2003). The Haldane malaria hypothesis: facts, artifacts, and a prophecy. *Redox Rep.* 8, 311–316. doi: 10.1179/135100003225002952

Amambua-Ngwa, A., Tetteh, K. K., Manske, M., Gomez-Escobar, N., Stewart, L. B., Deerhake, M. E., et al. (2012). Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet.* 8:e1002992. doi: 10.1371/journal.pgen.1002992

Amani, V., Boubou, M. I., Pied, S., Marussig, M., Walliker, D., Mazier, D., et al. (1998). Cloned lines of *Plasmodium berghei* ANKA differ in their abilities to induce experimental cerebral malaria. *Infect. Immun.* 66, 4093–4099. doi: 10.1128/IAI.66.9.4093-4099.1998

Ansari, H. R., Templeton, T. J., Subudhi, A. K., Ramaprasad, A., Tang, J., Lu, F., et al. (2016). Genome-scale comparison of expanded gene families in *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* with *Plasmodium malariae* and with other *Plasmodium* species. *Int. J. Parasitol.* 46, 685–696. doi: 10.1016/j.ijpara.2016.05.009

Asad, S., Feitosa-Suntheimer, F., Gold, A., Londono-Renteria, B., and Colpitts, T. M. (2019). Quantification of antibody-dependent enhancement of the Zika virus in primary human cells. *J. Vis. Exp.* 143, 1–6. doi: 10.3791/58691

Ashley, E. A., Pyae Phyo, A., and Woodrow, C. J. (2018). Malaria. *Lancet* 391, 1608–1621. doi: 10.1016/S0140-6736(18)30324-6

Auburn, S., Bohme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., et al. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res.* 1:4. doi: 10.12688/wellcomeopenres.9876.1

Avril, M., Brazier, A. J., Melcher, M., Sampath, S., and Smith, J. D. (2013). DC8 and DC13 var genes associated with severe malaria bind avidly to diverse endothelial cells. *PLoS Pathog.* 9:e1003430. doi: 10.1371/journal.ppat.1003430

Bachmann, A., Bruske, E., Krumkamp, R., Turner, L., Wichers, J. S., Petter, M., et al. (2019). Controlled human malaria infection with *Plasmodium falciparum* demonstrates impact of naturally acquired immunity on virulence gene expression. *PLoS Pathog.* 15:e1007906. doi: 10.1371/journal.ppat.1007906

Baer, K., Roosevelt, M., Clarkson, A. B. Jr., van Rooijen, N., Schnieder, T., and Frevert, U. (2007). Kupffer cells are obligatory for *Plasmodium yoelii* sporozoite infection of the liver. *Cell. Microbiol.* 9, 397–412. doi: 10.1111/j.1462-5822.2006.00798.x

Bando, H., Okado, K., Guelbeogo, W. M., Badolo, A., Aonuma, H., Nelson, B., et al. (2013). Intra-specific diversity of *Serratia marcescens* in *Anopheles* mosquito midgut defines *Plasmodium* transmission capacity. *Sci. Rep.* 3:1641. doi: 10.1038/srep01641

Barclay, V. C., Chan, B. H., Anders, R. F., and Read, A. F. (2008). Mixed allele malaria vaccines: host protection and within-host selection. *Vaccine* 26, 6099–6107. doi: 10.1016/j.vaccine.2008.09.004

Baruch, D. I., Pasloske, B. L., Singh, H. B., Bi, X., Ma, X. C., Feldman, M., et al. (1995). Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* 82, 77–87. doi: 10.1016/0092-8674(95)90054-3

Beeson, J. G., Kurtovic, L., Dobano, C., Opi, D. H., Chan, J. A., Feng, G., et al. (2019). Challenges and strategies for developing efficacious and long-lasting malaria vaccines. *Sci. Transl. Med.* 11:eaau1458. doi: 10.1126/scitranslmed.aau1458

Billi, A. C., Kahlenberg, J. M., and Gudjonsson, J. E. (2019). Sex bias in autoimmunity. *Curr. Opin. Rheumatol.* 31, 53–61. doi: 10.1097/BOR.0000000000000564

Biryukov, S., Angov, E., Landmesser, M. E., Spring, M. D., Ockenhouse, C. F., and Stoute, J. A. (2016). Complement and antibody-mediated enhancement of red blood cell invasion and growth of Malaria parasites. *EBioMedicine* 9, 207–216. doi: 10.1016/j.ebiom.2016.05.015

Blacklock, E., and Adler, S. (1922). A parasite resembling *Plasmodium falciparum* in chimpanzee. *Ann. Trop. Med. Prasitol.* 16, 99–107. doi: 10.1080/00034983.1922.11684303

Bohme, U., Otto, T. D., Cotton, J. A., Steinbiss, S., Sanders, M., Oyola, S. O., et al. (2018). Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals. *Genome Res.* 28, 547–560. doi: 10.1101/gr.218123.116

Boundenga, L., Ollomo, B., Rougeron, V., Mouele, L. Y., Mve-Ondo, B., Delicat-Loembet, L. M., et al. (2015). Diversity of malaria parasites in great apes in Gabon. *Malar. J.* 14:111. doi: 10.1186/s12936-015-0622-6

Bourgard, C., Albrecht, L., Kayano, A., Sunnerhagen, P., and Costa, F. T. M. (2018). *Plasmodium vivax* biology: insights provided by genomics, transcriptomics and proteomics. *Front. Cell. Infect. Microbiol.* 8:34. doi: 10.3389/fcimb.2018.00034

Brancucci, N. M. B., Gerdt, J. P., Wang, C., de Niz, M., Philip, N., Adapa, S. R., et al. (2017). Lysophosphatidylcholine regulates sexual stage differentiation in the human malaria parasite *Plasmodium falciparum*. *Cell* 171, 1532–1544.e1515. doi: 10.1016/j.cell.2017.10.020

Canali, S. (2008). Researches on thalassemia and malaria in Italy and the origins of the "Haldane hypothesis". *Med. Secoli.* 20, 827–846.

Carlton, J., Silva, J., and Hall, N. (2005). The genome of model malaria parasites, and comparative genomics. *Curr. Issues Mol. Biol.* 7, 23–37.

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763. doi: 10.1038/nature07327

Carter, R., and Mendis, K. N. (2002). Evolutionary and historical aspects of the burden of malaria. *Clin. Microbiol. Rev.* 15, 564–594. doi: 10.1128/CMR.15.4.564-594.2002

Chan, E. R., Barnwell, J. W., Zimmerman, P. A., and Serre, D. (2015). Comparative analysis of field-isolate and monkey-adapted *Plasmodium vivax* genomes. *PLoS Negl. Trop. Dis.* 9:e0003566. doi: 10.1371/journal.pntd.0003566

Cheesman, S., O'Mahony, E., Pattaradilokrat, S., Degnan, K., Knott, S., and Carter, R. (2010). A single parasite gene determines strain-specific protective immunity against malaria: the role of the merozoite surface protein I. *Int. J. Parasitol.* 40, 951–961. doi: 10.1016/j.ijpara.2010.02.003

Cheesman, S., Raza, A., and Carter, R. (2006). Mixed strain infections and strain-specific protective immunity in the rodent malaria parasite *Plasmodium chabaudi chabaudi* in mice. *Infect. Immun.* 74, 2996–3001. doi: 10.1128/IAI.74.5.2996-3001.2006

Chen, D. S., Barry, A. E., Leliwa-Sytek, A., Smith, T. A., Peterson, I., Brown, S. M., et al. (2011). A molecular epidemiological study of var gene diversity to characterize the reservoir of *Plasmodium falciparum* in humans in Africa. *PLoS ONE* 6:e16629. doi: 10.1371/journal.pone.0016629

Chen, Q., Fernandez, V., Sundstrom, A., Schlichtherle, M., Datta, S., Hagblom, P., et al. (1998). Developmental selection of var gene expression in *Plasmodium falciparum*. *Nature* 394, 392–395. doi: 10.1038/28660

Claessens, A., Hamilton, W. L., Kekre, M., Otto, T. D., Faizullabhoy, A., Rayner, J. C., et al. (2014). Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of Var genes during mitosis. *PLoS Genet.* 10:e1004812. doi: 10.1371/journal.pgen.1004812

Conway, D. J. (2015). Paths to a malaria vaccine illuminated by parasite genomics. *Trends Genet.* 31, 97–107. doi: 10.1016/j.tig.2014.12.005

Conway, D. J., Greenwood, B. M., and McBride, J. S. (1991). The epidemiology of multiple-clone *Plasmodium falciparum* infections in Gambian patients. *Parasitology* 103(Pt. 1), 1–6. doi: 10.1017/S0031182000059217

Cuesta-Astroz, Y., Santos, A., Oliveira, G., and Jensen, L. J. (2019). Analysis of predicted host-parasite interactomes reveals commonalities and specificities related to parasitic lifestyle and tissues tropism. *Front. Immunol.* 10:212. doi: 10.3389/fimmu.2019.00212

Culleton, R., and Kaneko, O. (2010). Erythrocyte binding ligands in malaria parasites: intracellular trafficking and parasite virulence. *Acta Trop.* 114, 131–137. doi: 10.1016/j.actatropica.2009.10.025

Culleton, R., Martinelli, A., Hunt, P., and Carter, R. (2005). Linkage group selection: rapid gene discovery in malaria parasites. *Genome Res.* 15, 92–97. doi: 10.1101/gr.2866205

Culleton, R. L., Inoue, M., Reece, S. E., Cheesman, S., and Carter, R. (2011). Strain-specific immunity induced by immunization with pre-erythrocytic stages of *Plasmodium chabaudi*. *Parasite Immunol.* 33, 73–78. doi: 10.1111/j.1365-3024.2010.01251.x

Cunningham, D., Fonager, J., Jarra, W., Carret, C., Preiser, P., and Langhorne, J. (2009). Rapid changes in transcription profiles of the *Plasmodium yoelii* yir multigene family in clonal populations: lack of epigenetic memory? *PLoS ONE* 4:e4285. doi: 10.1371/journal.pone.0004285

Cunningham, D., Lawton, J., Jarra, W., Preiser, P., and Langhorne, J. (2010). The pir multigene family of *Plasmodium*: antigenic variation and beyond. *Mol. Biochem. Parasitol.* 170, 65–73. doi: 10.1016/j.molbiopara.2009.12.010

Damena, D., Denis, A., Golassa, L., and Chimusa, E. R. (2019). Genome-wide association studies of severe *P. falciparum* malaria susceptibility: progress, pitfalls and prospects. *BMC Med Genomics* 12:120. doi: 10.1186/s12920-019-0564-x

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., et al. (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Mol. Biol. Evol.* 30, 1544–1558. doi: 10.1093/molbev/mst080

Day, K. P., Artzy-Randrup, Y., Tiedje, K. E., Rougeron, V., Chen, D. S., Rask, T. S., et al. (2017). Evidence of strain structure in *Plasmodium falciparum v*ar gene repertoires in children from Gabon, West Africa. *Proc. Natl. Acad. Sci. U.S.A.* 114, E4103–E4111. doi: 10.1073/pnas.1613018114

de Jong, R. M., Tebeje, S. K., Meerstein-Kessel, L., Tadesse, F. G., Jore, M. M., Stone, W., et al. (2020). Immunity against sexual stage *Plasmodium falciparum* and *Plasmodium vivax* parasites. *Immunol. Rev* 293, 190–215. doi: 10.1111/imr.12828

Deitsch, K. W., and Dzikowski, R. (2017). Variant gene expression and antigenic variation by malaria parasites. *Annu. Rev. Microbiol.* 71, 625–641. doi: 10.1146/annurev-micro-090816-093841

Dewasurendra, R. L., Jeffreys, A., Gunawardena, S. A., Chandrasekharan, N. V., Rockett, K., Kwiatkowski, D., et al. (2018). Host genetic polymorphisms and serological response against malaria in a selected population in Sri Lanka. *Malar. J.* 17:473. doi: 10.1186/s12936-018-2622-9

Draper, S. J., Sack, B. K., King, C. R., Nielsen, C. M., Rayner, J. C., Higgins, M. K., et al. (2018). Malaria vaccines: recent advances and new horizons. *Cell Host Microbe* 24, 43–56. doi: 10.1016/j.chom.2018.06.008

Drew, D. R., Hodder, A. N., Wilson, D. W., Foley, M., Mueller, I., Siba, P. M., et al. (2012). Defining the antigenic diversity of *Plasmodium falciparum* apical

membrane antigen 1 and the requirements for a multi-allele vaccine against malaria. *PLoS ONE* 7:e51023. doi: 10.1371/journal.pone.0051023

Duval, L., and Ariey, F. (2012). Ape *Plasmodium* parasites as a source of human outbreaks. *Clin. Microbiol. Infect.* 18, 528–532. doi: 10.1111/j.1469-0691.2012.03825.x

Duval, L., Fourment, M., Nerrienet, E., Rousset, D., Sadeuh, S. A., Goodman, S. M., et al. (2010). African apes as reservoirs of *Plasmodium falciparum* and the origin and diversification of the Laverania subgenus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10561–10566. doi: 10.1073/pnas.1005435107

Early, A. M., Lievens, M., MacInnis, B. L., Ockenhouse, C. F., Volkman, S. K., Adjei, S., et al. (2018). Host-mediated selection impacts the diversity of *Plasmodium falciparum* antigens within infections. *Nat. Commun.* 9:1381. doi: 10.1038/s41467-018-03807-7

Ebel, E. R., Telis, N., Venkataram, S., Petrov, D. A., and Enard, D. (2017). High rate of adaptation of mammalian proteins that interact with *Plasmodium* and related parasites. *PLoS Genet.* 13:e1007023. doi: 10.1371/journal.pgen.1007023

Escalante, A. A., Cornejo, O. E., Freeland, D. E., Poe, A. C., Durrego, E., Collins, W. E., et al. (2005). A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1980–1985. doi: 10.1073/pnas.0409652102

Escalante, A. A., and Pacheco, M. A. (2019). Malaria molecular epidemiology: an evolutionary genetics perspective. *Microbiol. Spectr.* 7, 1–17. doi: 10.1128/microbiolspec.AME-0010-2019

Fecchio, A., Chagas, C. R. F., Bell, J. A., and Kirchgatter, K. (2020). Evolutionary ecology, taxonomy, and systematics of avian malaria and related parasites. *Acta Trop.* 204:105364. doi: 10.1016/j.actatropica.2020.105364

Felger, I., Maire, M., Bretscher, M. T., Falk, N., Tiaden, A., Sama, W., et al. (2012). The dynamics of natural *Plasmodium falciparum* infections. *PLoS ONE* 7:e45542. doi: 10.1371/journal.pone.0045542

Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., et al. (2000). Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* 6, 861–871. doi: 10.1016/S1097-2765(05)00077-8

Fish, E. N. (2008). The X-files in immunity: sex-based differences predispose immune responses. *Nat. Rev. Immunol.* 8, 737–744. doi: 10.1038/nri2394

Fluck, C., Smith, T., Beck, H. P., Irion, A., Betuela, I., Alpers, M. P., et al. (2004). Strain-specific humoral response to a polymorphic malaria vaccine. *Infect. Immun.* 72, 6300–6305. doi: 10.1128/IAI.72.11.6300-6305.2004

Frank, M., Kirkman, L., Costantini, D., Sanyal, S., Lavazec, C., Templeton, T. J., et al. (2008). Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. *Int. J. Parasitol.* 38, 1099–1109. doi: 10.1016/j.ijpara.2008.01.010

Frimpong, A., Kusi, K. A., Ofori, M. F., and Ndifon, W. (2018). Novel strategies for malaria vaccine design. *Front. Immunol.* 9:2769. doi: 10.3389/fimmu.2018.02769

Galaway, F., Yu, R., Constantinou, A., Prugnolle, F., and Wright, G. J. (2019). Resurrection of the ancestral RH5 invasion ligand provides a molecular explanation for the origin of *P. falciparum* malaria in humans. *PLOS Biol.* 17:e3000490. doi: 10.1371/journal.pbio.3000490

Galen, S. C., Borner, J., Martinsen, E. S., Schaer, J., Austin, C. C., West, C. J., et al. (2018). The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. *R Soc. Open Sci.* 5:171780. doi: 10.1098/rsos.171780

Galinski, M. R. (2019). Functional genomics of simian malaria parasites and host-parasite interactions. *Brief. Funct. Genomics* 18, 270–280. doi: 10.1093/bfgp/elz013

Galinski, M. R., Lapp, S. A., Peterson, M. S., Ay, F., Joyner, C. J., KG, L.E. R., et al. (2018). *Plasmodium knowlesi*: a superb *in vivo* nonhuman primate model of antigenic variation in malaria. *Parasitology* 145, 85–100. doi: 10.1017/S0031182017001135

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511. doi: 10.1038/nature01097

Garrido-Cardenas, J. A., Gonzalez-Ceron, L., Manzano-Agugliaro, F., and Mesa-Valle, C. (2019). *Plasmodium* genomics: an approach for learning about and ending human malaria. *Parasitol. Res.* 118, 1–27. doi: 10.1007/s00436-018-6127-9

Georgiadou, A., Lee, H. J., Walther, M., van Beek, A. E., Fitriani, F., Wouters, D., et al. (2019). Modelling pathogen load dynamics to elucidate mechanistic determinants of host-*Plasmodium falciparum* interactions. *Nat. Microbiol.* 4, 1592–1602. doi: 10.1038/s41564-019-0474-x

Goheen, M. M., Campino, S., and Cerami, C. (2017). The role of the red blood cell in host defence against *falciparum* malaria: an expanding repertoire of evolutionary alterations. *Br. J. Haematol.* 179, 543–556. doi: 10.1111/bjh.14886

Grajcarek, J., Monlong, J., Nishinaka-Arai, Y., Nakamura, M., Nagai, M., Matsuo, S., et al. (2019). Genome-wide microhomologies enable precise template-free editing of biologically relevant deletion mutations. *Nat. Commun.* 10:4856. doi: 10.1038/s41467-019-12829-8

Graves, S. F., Kouriba, B., Diarra, I., Daou, M., Niangaly, A., Coulibaly, D., et al. (2016). Strain-specific *Plasmodium falciparum* multifunctional CD4(+) T cell cytokine expression in Malian children immunized with the FMP2.1/AS02A vaccine candidate. *Vaccine* 34, 2546–2555. doi: 10.1016/j.vaccine.2016.04.019

Greenwood, J. M., Ezquerra, A. L., Behrens, S., Branca, A., and Mallet, L. (2016). Current analysis of host-parasite interactions with a focus on next generation sequencing data. *Zoology* 119, 298–306. doi: 10.1016/j.zool.2016.06.010

Halstead, S. B., Mahalingam, S., Marovich, M. A., Ubol, S., and Mosser, D. M. (2010). Intrinsic antibody-dependent enhancement of microbial infection in macrophages: disease regulation by immune complexes. *Lancet Infect. Dis.* 10, 712–722. doi: 10.1016/S1473-3099(10)70166-3

Hedrick, P. W. (2012). Resistance to malaria in humans: the impact of strong, recent selection. *Malar. J.* 11:349. doi: 10.1186/1475-2875-11-349

Howard, R. J., Barnwell, J. W., Rock, E. P., Neequaye, J., Ofori-Adjei, D., Maloy, W. L., et al. (1988). Two approximately 300 kilodalton *Plasmodium falciparum* proteins at the surface membrane of infected erythrocytes. *Mol. Biochem. Parasitol.* 27, 207–223. doi: 10.1016/0166-6851(88)90040-0

Hviid, L., and Jensen, A. T. (2015). PfEMP1 - a parasite protein family of key importance in *Plasmodium falciparum* Malaria immunity and pathogenesis. *Adv. Parasitol.* 88, 51–84. doi: 10.1016/bs.apar.2015.02.004

Idaghdour, Y., Quinlan, J., Goulet, J. P., Berghout, J., Gbeha, E., Bruat, V., et al. (2012). Evidence for additive and interaction effects of host genotype and infection in malaria. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16786–16793. doi: 10.1073/pnas.1204945109

Janssen, C. S., Barrett, M. P., Turner, C. M., and Phillips, R. S. (2002). A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proc. Biol. Sci.* 269, 431–436. doi: 10.1098/rspb.2001.1903

Jeffares, D. C., Pain, A., Berry, A., Cox, A. V., Stalker, J., Ingle, C. E., et al. (2007). Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat. Genet.* 39, 120–125. doi: 10.1038/ng1931

Jemmely, N. Y., Niang, M., and Preiser, P. R. (2010). Small variant surface antigens and *Plasmodium* evasion of immunity. *Fut. Microbiol.* 5, 663–682. doi: 10.2217/fmb.10.21

Jensen, A. R., Adams, Y., and Hviid, L. (2020). Cerebral *Plasmodium falciparum* malaria: the role of PfEMP1 in its pathogenesis and immunity, and PfEMP1-based vaccines to prevent it. *Immunol. Rev.* 293, 230–252. doi: 10.1111/imr.12807

Josling, G. A., and Llinas, M. (2015). Sexual development in *Plasmodium* parasites: knowing when it's time to commit. *Nat. Rev. Microbiol.* 13, 573–587. doi: 10.1038/nrmicro3519

Joy, D. A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A. U., et al. (2003). Early origin and recent expansion of *Plasmodium falciparum*. *Science* 300, 318–321. doi: 10.1126/science.1081449

Joy, D. A., Gonzalez-Ceron, L., Carlton, J. M., Gueye, A., Fay, M., McCutchan, T. F., et al. (2008). Local adaptation and vector-mediated population structure in *Plasmodium vivax malaria*. *Mol. Biol. Evol.* 25, 1245–1252. doi: 10.1093/molbev/msn073

Kariuki, S. N., and Williams, T. N. (2020). Human genetics and malaria resistance. *Hum. Genet.* 139, 801–811. doi: 10.1007/s00439-020-02142-6

Kirchner, S., Power, B. J., and Waters, A. P. (2016). Recent advances in malaria genomics and epigenomics. *Genome Med.* 8:92. doi: 10.1186/s13073-016-0343-7

Kirkman, L., and Deitsch, K. W. (2020). Vive la différence: exploiting the differences between rodent and human malarias. *Trends Parasitol.* 36, 504–511. doi: 10.1016/j.pt.2020.03.008

Kirkman, L. A., Lawrence, E. A., and Deitsch, K. W. (2014). Malaria parasites utilize both homologous recombination and alternative end joining

pathways to maintain genome integrity. *Nucleic Acids Res.* 42, 370–379. doi: 10.1093/nar/gkt881

Klein, S. L., Shann, F., Moss, W. J., Benn, C. S., and Aaby, P. (2016). RTS,S malaria vaccine and increased mortality in girls. *MBio* 7:e00514–00516. doi: 10.1128/mBio.00514-16

Kobbe, R., Neuhoff, R., Marks, F., Adjei, S., Langefeld, I., von Reden, C., et al. (2006). Seasonal variation and high multiplicity of first *Plasmodium falciparum* infections in children from a holoendemic area in Ghana, West Africa. *Trop. Med. Int. Health* 11, 613–619. doi: 10.1111/j.1365-3156.2006.01618.x

Kooij, T. W., Carlton, J. M., Bidwell, S. L., Hall, N., Ramesar, J., Janse, C. J., et al. (2005). A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog.* 1:e44. doi: 10.1371/journal.ppat.0010044

Kotepui, M., Kotepui, K. U., De Jesus Milanez, G., and Masangkay, F. R. (2020). *Plasmodium* spp. mixed infection leading to severe malaria: a systematic review and meta-analysis. *Sci. Rep.* 10:11068. doi: 10.1038/s41598-020-68082-3

Krief, S., Escalante, A. A., Pacheco, M. A., Mugisha, L., André, C., Halbwax, M., et al. (2010). On the diversity of malaria parasites in African apes and the origin of *Plasmodium falciparum* from bonobos. *PLoS Pathog.* 6:e1000765. doi: 10.1371/journal.ppat.1000765

Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* 77, 171–192. doi: 10.1086/432519

Kwon, H., and Smith, R. C. (2019). Chemical depletion of phagocytic immune cells in *Anopheles gambiae* reveals dual roles of mosquito hemocytes in anti-*Plasmodium* immunity. *Proc. Natl. Acad. Sci. U.S.A.* 116, 14119–14128. doi: 10.1073/pnas.1900147116

Lacerda-Queiroz, N., Lima, O. C., Carneiro, C. M., Vilela, M. C., Teixeira, A. L., Teixeira-Carvalho, A., et al. (2011). *Plasmodium berghei* NK65 induces cerebral leukocyte recruitment *in vivo*: an intravital microscopic study. *Acta Trop.* 120, 31–39. doi: 10.1016/j.actatropica.2011.04.020

Lacerda-Queiroz, N., Riteau, N., Eastman, R. T., Bock, K. W., Orandle, M. S., Moore, I. N., et al. (2017). Mechanism of splenic cell death and host mortality in a *Plasmodium yoelii* malaria model. *Sci. Rep.* 7:10438. doi: 10.1038/s41598-017-10776-2

Larsen, M. D., Quintana, M. D. P., Ditlev, S. B., Bayarri-Olmos, R., Ofori, M. F., Hviid, L., et al. (2018). Evasion of classical complement pathway activation on *Plasmodium falciparum*-infected erythrocytes opsonized by PfEMP1-specific IgG. *Front. Immunol.* 9:3088. doi: 10.3389/fimmu.2018.03088

Laurens, M. B. (2018). The promise of a malaria vaccine-are we closer? *Annu. Rev. Microbiol.* 72, 273–292. doi: 10.1146/annurev-micro-090817-062427

Laurens, M. B., Kouriba, B., Bergmann-Leitner, E., Angov, E., Coulibaly, D., Diarra, I., et al. (2017). Strain-specific *Plasmodium falciparum* growth inhibition among Malian children immunized with a blood-stage malaria vaccine. *PLoS ONE* 12:e0173294. doi: 10.1371/journal.pone.0173294

Lauron, E. J., Aw Yeang, H. X., Taffner, S. M., and Sehgal, R. N. (2015). *De novo* assembly and transcriptome analysis of *Plasmodium gallinaceum* identifies the Rh5 interacting protein (ripr), and reveals a lack of EBL and RH gene family diversification. *Malar. J.* 14:296. doi: 10.1186/s12936-015-0814-0

Lawton, J., Brugat, T., Yan, Y. X., Reid, A. J., Bohme, U., Otto, T. D., et al. (2012). Characterization and gene expression analysis of the cir multi-gene family of *Plasmodium chabaudi chabaudi* (AS). *BMC Genomics* 13:125. doi: 10.1186/1471-2164-13-125

Lee, A. H., Symington, L. S., and Fidock, D. A. (2014). DNA repair mechanisms and their biological roles in the malaria parasite *Plasmodium falciparum*. *Microbiol. Mol. Biol. Rev.* 78, 469–486. doi: 10.1128/MMBR.00059-13

Lee, H. J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L. B., Levin, M., et al. (2018). Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Sci. Transl. Med.* 10:eaar3619. doi: 10.1126/scitranslmed.aar3619

Liu, W., Li, Y., Learn, G. H., Rudicell, R. S., Robertson, J. D., Keele, B. F., et al. (2010). Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* 467, 420–425. doi: 10.1038/nature09442

Liu, W., Sherrill-Mix, S., Learn, G. H., Scully, E. J., Li, Y., Avitto, A. N., et al. (2017). Wild bonobos host geographically restricted malaria parasites including a putative new *Laverania* species. *Nat. Commun.* 8:1635. doi: 10.1038/s41467-017-01798-5

Lover, A. A., and Coker, R. J. (2015). Do mixed infections matter? Assessing virulence of mixed-clone infections in experimental human and murine malaria. *Infect. Genet. Evol.* 36, 82–91. doi: 10.1016/j.meegid.2015.08.041

Luzolo, A. L., and Ngoyi, D. M. (2019). Cerebral malaria. *Brain Res. Bull.* 145, 53–58. doi: 10.1016/j.brainresbull.2019.01.010

Mahanta, A., Ganguli, P., Barah, P., Sarkar, R. R., Sarmah, N., Phukan, S., et al. (2018). Integrative approaches to understand the mastery in manipulation of host cytokine networks by protozoan parasites with emphasis on *Plasmodium* and *Leishmania* species. *Front. Immunol.* 9:296. doi: 10.3389/fimmu.2018.00296

Malaria Genomic Epidemiology Network. (2019). Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.* 10:5732. doi: 10.1038/s41467-019-13480-z

Mangano, V. D., and Modiano, D. (2014). An evolutionary perspective of how infection drives human genome diversity: the case of malaria. *Curr. Opin. Immunol.* 30, 39–47. doi: 10.1016/j.coi.2014.06.004

Martin, M. J., Rayner, J. C., Gagneux, P., Barnwell, J. W., and Varki, A. (2005). Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of N-glycolylneuraminic acid. *Proc. Natl. Acad. Sci. U. S.A.* 102, 12819–12824. doi: 10.1073/pnas.0503819102

Martinelli, A., and Culleton, R. (2018). Non-human primate malaria parasites: out of the forest and into the laboratory. *Parasitology* 145, 41–54. doi: 10.1017/S0031182016001335

McCollum, A. M., Basco, L. K., Tahar, R., Udhayakumar, V., and Escalante, A. A. (2008). Hitchhiking and selective sweeps of *Plasmodium falciparum* sulfadoxine and pyrimethamine resistance alleles in a population from central Africa. *Antimicrob. Agents Chemother.* 52, 4089–4097. doi: 10.1128/AAC.00623-08

Mejias, A., Suarez, N. M., and Ramilo, O. (2014). Detecting specific infections in children through host responses: a paradigm shift. *Curr. Opin. Infect. Dis.* 27, 228–235. doi: 10.1097/QCO.0000000000000065

Molina-Cruz, A., Canepa, G. E., Kamath, N., Pavlovic, N. V., Mu, J., Ramphul, U. N., et al. (2015). *Plasmodium* evasion of mosquito immunity and global malaria transmission: The lock-and-key theory. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15178–15183. doi: 10.1073/pnas.1520426112

Mu, J., Awadalla, P., Duan, J., McGee, K. M., Keebler, J., Seydel, K., et al. (2007). Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat. Genet.* 39, 126–130. doi: 10.1038/ng1924

Mu, J., Duan, J., Makova, K. D., Joy, D. A., Huynh, C. Q., Branch, O. H., et al. (2002). Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* 418, 323–326. doi: 10.1038/nature00836

Muller, O., Tozan, Y., and Becher, H. (2015). RTS,S/AS01 malaria vaccine and child mortality. *Lancet* 386:1736. doi: 10.1016/S0140-6736(15)00694-7

Nacer, A., Movila, A., Sohet, F., Girgis, N. M., Gundra, U. M., Loke, P., et al. (2014). Experimental cerebral malaria pathogenesis–hemodynamics at the blood brain barrier. *PLoS Pathog.* 10:e1004528. doi: 10.1371/journal.ppat.1004528

Neafsey, D. E., Hartl, D. L., and Berriman, M. (2005). Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. *Mol. Biol. Evol.* 22, 1621–1626. doi: 10.1093/molbev/msi154

Neafsey, D. E., Juraska, M., Bedford, T., Benkeser, D., Valim, C., Griggs, A., et al. (2015). Genetic diversity and protective efficacy of the RTS,S/AS01 malaria vaccine. *N. Engl. J. Med.* 373, 2025–2037. doi: 10.1056/NEJMoa1505819

Negro, F. (2020). Is antibody-dependent enhancement playing a role in COVID-19 pathogenesis? *Swiss Med. Wkly.* 150:w20249. doi: 10.4414/smw.2020.20249

O'Brien, J. D., Iqbal, Z., Wendler, J., and Amenga-Etego, L. (2016). Inferring strain mixture within clinical *Plasmodium falciparum* isolates from genomic sequence data. *PLoS Comput. Biol.* 12:e1004824. doi: 10.1371/journal.pcbi.1004824

Ochola, L. I., Tetteh, K. K., Stewart, L. B., Riitho, V., Marsh, K., and Conway, D. J. (2010). Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Mol. Biol. Evol.* 27, 2344–2351. doi: 10.1093/molbev/msq119

Ollomo, B., Durand, P., Prugnolle, F., Douzery, E., Arnathau, C., Nkoghe, D., et al. (2009). A new malaria agent in African hominids. *PLoS Pathog.* 5:e1000446. doi: 10.1371/journal.ppat.1000446

Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Leach, A., Lievens, M., et al. (2016). Seven-year efficacy of RTS,S/AS01 malaria vaccine among young African children. *N. Engl. J. Med.* 374, 2519–2529. doi: 10.1056/NEJMoa1515257

Otsuki, H., Kaneko, O., Thongkukiatkul, A., Tachibana, M., Iriko, H., Takeo, S., et al. (2009). Single amino acid substitution in *Plasmodium yoelii* erythrocyte ligand determines its localization and controls parasite virulence. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7167–7172. doi: 10.1073/pnas.0811313106

Otto, T. D., Bohme, U., Jackson, A. P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W. A., et al. (2014a). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol.* 12:86. doi: 10.1186/s12915-014-0086-0

Otto, T. D., Gilabert, A., Crellen, T., Böhme, U., Arnathau, C., Sanders, M., et al. (2018). Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* 3, 687–697. doi: 10.1038/s41564-018-0162-2

Otto, T. D., Rayner, J. C., Bohme, U., Pain, A., Spottiswoode, N., Sanders, M., et al. (2014b). Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat. Commun.* 5:4754. doi: 10.1038/ncomms5754

Ouattara, A., Mu, J., Takala-Harrison, S., Saye, R., Sagara, I., Dicko, A., et al. (2010). Lack of allele-specific efficacy of a bivalent AMA1 malaria vaccine. *Malar. J.* 9:175. doi: 10.1186/1475-2875-9-175

Pain, A., Bohme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., et al. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455, 799–803. doi: 10.1038/nature07306

Pasini, E. M., Bohme, U., Rutledge, G. G., Voorberg-Van der Wel, A., Sanders, M., Berriman, M., et al. (2017). An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Res.* 2:42. doi: 10.12688/wellcomeopenres.11864.1

Pattaradilokrat, S., Cheesman, S. J., and Carter, R. (2008). Congenicity and genetic polymorphism in cloned lines derived from a single isolate of a rodent malaria parasite. *Mol. Biochem. Parasitol.* 157, 244–247. doi: 10.1016/j.molbiopara.2007.10.011

Pattaradilokrat, S., Culleton, R. L., Cheesman, S. J., and Carter, R. (2009). Gene encoding erythrocyte binding ligand linked to blood stage multiplication rate phenotype in *Plasmodium yoelii yoelii*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7161–7166. doi: 10.1073/pnas.0811430106

Peng, Y. C., Qi, Y., Zhang, C., Yao, X., Wu, J., Pattaradilokrat, S., et al. (2020). *Plasmodium yoelii* erythrocyte-binding-like protein modulates host cell membrane structure, immunity, and disease severity. *MBio* 11:e02995-19. doi: 10.1128/mBio.02995-19

Perkins, S. L., Sarkar, I. N., and Carter, R. (2007). The phylogeny of rodent malaria parasites: simultaneous analysis across three genomes. *Infect. Genet. Evol.* 7, 74–83. doi: 10.1016/j.meegid.2006.04.005

Peters, J., Fowler, E., Gatton, M., Chen, N., Saul, A., and Cheng, Q. (2002). High diversity and rapid changeover of expressed var genes during the acute phase of *Plasmodium falciparum* infections in human volunteers. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10689–10694. doi: 10.1073/pnas.162349899

Peterson, D. S., Milhous, W. K., and Wellems, T. E. (1990). Molecular basis of differential resistance to cycloguanil and pyrimethamine in *Plasmodium falciparum* malaria. *Proc. Natl. Acad. Sci. U.S.A.* 87, 3018–3022. doi: 10.1073/pnas.87.8.3018

Plenderleith, L. J., Liu, W., Learn, G. H., Loy, D. E., Speede, S., Sanz, C. M., et al. (2019). Ancient introgression between two ape malaria parasite species. *Genome Biol. Evol.* 11, 3269–3274. doi: 10.1093/gbe/evz244

Porterfield, J. S. (1986). Antibody-dependent enhancement of viral infectivity. *Adv. Virus Res.* 31, 335–355. doi: 10.1016/S0065-3527(08)60268-7

Proto, W. R., Siegel, S. V., Dankwa, S., Liu, W., Kemp, A., Marsden, S., et al. (2019). Adaptation of *Plasmodium falciparum* to humans involved the loss of an ape-specific erythrocyte invasion ligand. *Nat. Commun.* 10:4512. doi: 10.1038/s41467-019-12294-3

Prugnolle, F., Durand, P., Neel, C., Ollomo, B., Ayala, F. J., Arnathau, C., et al. (2010). African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1458–1463. doi: 10.1073/pnas.0914440107

Quirk, S. J., Haraldsson, G., Erlendsdottir, H., Hjalmarsdottir, M. A., van Tonder, A. J., Hrafnkelsson, B., et al. (2018). Effect of vaccination on *Pneumococci* isolated from the nasopharynx of healthy children and the middle ear of children with otitis media in Iceland. *J. Clin. Microbiol.* 56:e01046-18. doi: 10.1128/JCM.01046-18

Ramiro, R. S., Reece, S. E., and Obbard, D. J. (2012). Molecular evolution and phylogenetics of rodent malaria parasites. *BMC Evol. Biol.* 12:219. doi: 10.1186/1471-2148-12-219

Ranjit, M. R., and Sharma, Y. D. (1999). Genetic polymorphism of *falciparum* malaria vaccine candidate antigen genes among field isolates in India. *Am. J. Trop. Med. Hyg.* 61, 103–108. doi: 10.4269/ajtmh.1999.61.103

Rask, T. S., Hansen, D. A., Theander, T. G., Gorm Pedersen, A., and Lavstsen, T. (2010). *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. *PLoS Comput. Biol.* 6:e1000933. doi: 10.1371/journal.pcbi.1000933

Ravenhall, M., Campino, S., Sepulveda, N., Manjurano, A., Nadjm, B., Mtove, G., et al. (2018). Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLoS Genet.* 14:e1007172. doi: 10.1371/journal.pgen.1007172

Rayner, J. C., Liu, W., Peeters, M., Sharp, P. M., and Hahn, B. H. (2011). A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends Parasitol.* 27, 222–229. doi: 10.1016/j.pt.2011.01.006

Recker, M., Buckee, C. O., Serazin, A., Kyes, S., Pinches, R., Christodoulou, Z., et al. (2011). Antigenic variation in *Plasmodium falciparum* malaria involves a highly structured switching pattern. *PLoS Pathog.* 7:e1001306. doi: 10.1371/journal.ppat.1001306

Reichenow, E. (1920). Uber das vorkommen der malaiaparasiten des menschen bei den Afrikanischen menschenaffen. *Ventralbl. Bakteriol. Parasitenkd. Intekionskr. Hyg. Abt.* 85, 207–216.

Reid, A. J. (2015). Large, rapidly evolving gene families are at the forefront of host-parasite interactions in *Apicomplexa*. *Parasitology* 142(Suppl. 1), S57–S70. doi: 10.1017/S0031182014001528

Rich, S. M., Leendertz, F. H., Xu, G., LeBreton, M., Djoko, C. F., Aminake, M. N., et al. (2009). The origin of malignant malaria. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14902–14907. doi: 10.1073/pnas.0907740106

Rich, S. M., Licht, M. C., Hudson, R. R., and Ayala, F. J. (1998). Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4425–4430. doi: 10.1073/pnas.95.8.4425

Riley, E. M. (1999). Is T-cell priming required for initiation of pathology in malaria infections? *Immunol. Today* 20, 228–233. doi: 10.1016/S0167-5699(99)01456-5

Roper, C., Pearce, R., Nair, S., Sharp, B., Nosten, F., and Anderson, T. (2004). Intercontinental spread of pyrimethamine-resistant malaria. *Science* 305:1124. doi: 10.1126/science.1098876

Rorick, M. M., Rask, T. S., Baskerville, E. B., Day, K. P., and Pascual, M. (2013). Homology blocks of *Plasmodium falciparum* var genes and clinically distinct forms of severe malaria in a local population. *BMC Microbiol.* 13:244. doi: 10.1186/1471-2180-13-244

Rts, S. C. T. P. (2015). Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* 386, 31–45. doi: 10.1016/S0140-6736(15)60721-8

Rts, S. C. T. P., Agnandji, S. T., Lell, B., Fernandes, J. F., Abossolo, B. P., Methogo, B. G., et al. (2012). A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *N. Engl. J. Med.* 367, 2284–2295. doi: 10.1056/NEJMoa1208394

Rts, S. C. T. P., Agnandji, S. T., Lell, B., Soulanoudjingar, S. S., Fernandes, J. F., Abossolo, B. P., et al. (2011). First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *N. Engl. J. Med.* 365, 1863–1875. doi: 10.1056/NEJMoa1102287

Rutledge, G. G., Bohme, U., Sanders, M., Reid, A. J., Cotton, J. A., Maiga-Ascofare, O., et al. (2017). *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature* 542, 101–104. doi: 10.1038/nature21038

Ruybal-Pesantez, S., Tiedje, K. E., Tonkin-Hill, G., Rask, T. S., Kamya, M. R., Greenhouse, B., et al. (2017). Population genomics of virulence genes of *Plasmodium falciparum* in clinical isolates from Uganda. *Sci. Rep.* 7:11810. doi: 10.1038/s41598-017-11814-9

Salinas, N. D., Tang, W. K., and Tolia, N. H. (2019). Blood-stage malaria parasite antigens: structure, function, and vaccine potential. *J. Mol. Biol.* 431, 4259–4280. doi: 10.1016/j.jmb.2019.05.018

Schaer, J., Perkins, S. L., Decher, J., Leendertz, F. H., Fahr, J., Weber, N., et al. (2013). High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17415–17419. doi: 10.1073/pnas.1311016110

Schall, J. J. (1982). Lizards infected with malaria: physiological and behavioral consequences. *Science* 217, 1057–1059. doi: 10.1126/science.7112113

Shutler, D., Reece, S. E., Mullie, A., Billingsley, P. F., and Read, A. F. (2005). Rodent malaria parasites *Plasmodium chabaudi* and *P. vinckei* do not increase their rates of gametocytogenesis in response to mosquito probing. *Proc. Biol. Sci.* 272, 2397–2402. doi: 10.1098/rspb.2005.3232

Siao, M. C., Borner, J., Perkins, S. L., Deitsch, K. W., and Kirkman, L. A. (2020). Evolution of host specificity by malaria parasites through altered mechanisms controlling genome maintenance. *mBio* 11:e03272-19. doi: 10.1128/mBio.03272-19

Smith, J. D., Chitnis, C. E., Craig, A. G., Roberts, D. J., Hudson-Taylor, D. E., Peterson, D. S., et al. (1995). Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* 82, 101–110. doi: 10.1016/0092-8674(95)90056-X

Smith, J. D., Rowe, J. A., Higgins, M. K., and Lavstsen, T. (2013). Malaria's deadly grip: cytoadhesion of *Plasmodium falciparum*-infected erythrocytes. *Cell. Microbiol.* 15, 1976–1983. doi: 10.1111/cmi.12183

Spence, P. J., Jarra, W., Levy, P., Reid, A. J., Chappell, L., Brugat, T., et al. (2013). Vector transmission regulates immune control of *Plasmodium* virulence. *Nature* 498, 228–231. doi: 10.1038/nature12231

Stone, W., Bousema, T., Sauerwein, R., and Drakeley, C. (2019). Two-faced immunity? The evidence for antibody enhancement of malaria transmission. *Trends Parasitol.* 35, 140–153. doi: 10.1016/j.pt.2018.11.003

Stone, W. J. R., Campo, J. J., Ouedraogo, A. L., Meerstein-Kessel, L., Morlais, I., Da, D., et al. (2018). Unravelling the immune signature of *Plasmodium falciparum* transmission-reducing immunity. *Nat. Commun.* 9:558. doi: 10.1038/s41467-017-02646-2

Su, X.-z., Mu, J., and Joy, D. A. (2003). The "Malaria's Eve" hypothesis and the debate concerning the origin of the human malaria parasite *Plasmodium falciparum*. *Microbes Infect.* 5, 891–896. doi: 10.1016/S1286-4579(03)00173-4

Su, X. Z., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., et al. (1995). The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* 82, 89–100. doi: 10.1016/0092-8674(95)90055-1

Su, X. Z., Lane, K. D., Xia, L., Sa, J. M., and Wellems, T. E. (2019). *Plasmodium* genomics and genetics: new insights into malaria pathogenesis, drug resistance, epidemiology, and evolution. *Clin. Microbiol. Rev.* 32:e00019-19. doi: 10.1128/CMR.00019-19

Sundararaman, S. A., Plenderleith, L. J., Liu, W., Loy, D. E., Learn, G. H., Li, Y., et al. (2016). Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat. Commun.* 7:11078. doi: 10.1038/ncomms11078

Sutherland, C. J., Tanomsing, N., Nolder, D., Oguike, M., Jennison, C., Pukrittayakamee, S., et al. (2010). Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium ovale* occur globally. *J. Infect. Dis.* 201, 1544–1550. doi: 10.1086/652240

Swardson-Olver, C. J., Dawson, T. C., Burnett, R. C., Peiper, S. C., Maeda, N., and Avery, A. C. (2002). *Plasmodium yoelii* uses the murine Duffy antigen receptor for chemokines as a receptor for normocyte invasion and an alternative receptor for reticulocyte invasion. *Blood* 99, 2677–2684. doi: 10.1182/blood.V99.8.2677

Sypniewska, P., Duda, J. F., Locatelli, I., Althaus, C. R., Althaus, F., and Genton, B. (2017). Clinical and laboratory predictors of death in African children with features of severe malaria: a systematic review and meta-analysis. *BMC Med.* 15:147. doi: 10.1186/s12916-017-0906-5

Tanabe, K., Zollner, G., Vaughan, J. A., Sattabongkot, J., Khuntirat, B., Honma, H., et al. (2015). *Plasmodium falciparum*: genetic diversity and complexity of infections in an isolated village in western Thailand. *Parasitol. Int.* 64, 260–266. doi: 10.1016/j.parint.2013.09.011

Tang, Y., Gupta, A., Garimalla, S., Ma, H. C. E., Galinski, M. R., Styczynski, M. P., et al. (2018). Metabolic modeling helps interpret transcriptomic changes during malaria. *Biochim. Biophys. Acta Mol. Basis Dis. 1864*, 2329–2340. doi: 10.1016/j.bbadis.2017.10.023

Tavares, J., Formaglio, P., Thiberge, S., Mordelet, E., van Rooijen, N., Medvinsky, A., et al. (2013). Role of host cell traversal by the malaria sporozoite during liver infection. *J. Exp. Med.* 210, 905–915. doi: 10.1084/jem.20121130

Templeton, T. J., Asada, M., Jiratanh, M., Ishikawa, S. A., Tiawsirisup, S., Sivakumar, T., et al. (2016). Ungulate malaria parasites. *Sci. Rep.* 6:23230. doi: 10.1038/srep23230

Tetteh, K. K., Stewart, L. B., Ochola, L. I., Amambua-Ngwa, A., Thomas, A. W., Marsh, K., et al. (2009). Prospective identification of malaria parasite genes under balancing selection. *PLoS ONE* 4:e5568. doi: 10.1371/journal.pone.0005568

Thiam, A., Sanka, M., Ndiaye Diallo, R., Torres, M., Mbengue, B., Nunez, N. F., et al. (2019). Gene expression profiling in blood from cerebral malaria patients and mild malaria patients living in Senegal. *BMC Med. Genomics* 12:148. doi: 10.1186/s12920-019-0599-z

Tishkoff, S. A., and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3, 611–621. doi: 10.1038/nrg865

Tonkin-Hill, G. Q., Trianty, L., Noviyanti, R., Nguyen, H. H. T., Sebayang, B. F., Lampah, D. A., et al. (2018). The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen-encoding var genes. *PLoS Biol.* 16:e2004328. doi: 10.1371/journal.pbio.2004328

Trampuz, A., Jereb, M., Muzlovic, I., and Prabhu, R. M. (2003). Clinical review: severe malaria. *Crit Care* 7, 315–323. doi: 10.1186/cc2183

Trimnell, A. R., Kraemer, S. M., Mukherjee, S., Phippard, D. J., Janes, J. H., Flamoe, E., et al. (2006). Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Mol. Biochem. Parasitol.* 148, 169–180. doi: 10.1016/j.molbiopara.2006.03.012

Verra, F., Mangano, V. D., and Modiano, D. (2009). Genetics of susceptibility to *Plasmodium falciparum*: from classical malaria resistance genes towards genome-wide association studies. *Parasite Immunol.* 31, 234–253. doi: 10.1111/j.1365-3024.2009.01106.x

Videvall, E. (2019). Genomic advances in avian malaria research. *Trends Parasitol.* 35, 254–266. doi: 10.1016/j.pt.2018.12.005

Vijayan, A., and Chitnis, C. E. (2019). Development of blood stage malaria vaccines. *Methods Mol. Biol.* 2013, 199–218. doi: 10.1007/978-1-4939-9550-9_15

Volkman, S. K., Barry, A. E., Lyons, E. J., Nielsen, K. M., Thomas, S. M., Choi, M., et al. (2001). Recent origin of *Plasmodium falciparum* from a single progenitor. *Science* 293, 482–484. doi: 10.1126/science.1059878

Volkman, S. K., Sabeti, P. C., deCaprio, D., Neafsey, D. E., Schaffner, S. F., Milner, D. A., et al. (2007). A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.* 39, 113–119. doi: 10.1038/ng1930

Waggoner, J. J., Katzelnick, L. C., Burger-Calderon, R., Gallini, J., Moore, R. H., Kuan, G., et al. (2020). Antibody-dependent enhancement of severe disease is mediated by serum viral load in pediatric dengue virus infections. *J. Infect. Dis.* 6:29201. doi: 10.1093/infdis/jiz618

Wahlgren, M., Goel, S., and Akhouri, R. R. (2017). Variant surface antigens of *Plasmodium falciparum* and their roles in severe malaria. *Nat. Rev. Microbiol* 15, 479–491. doi: 10.1038/nrmicro.2017.47

Wan, Y., Shang, J., Sun, S., Tai, W., Chen, J., Geng, Q., et al. (2020). Molecular mechanism for antibody-dependent enhancement of coronavirus entry. *J. Virol.* 94:e02015-19. doi: 10.1128/JVI.02015-19

Wanaguru, M., Liu, W., Hahn, B. H., Rayner, J. C., and Wright, G. J. (2013). RH5-Basigin interaction plays a major role in the host tropism of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20735–20740. doi: 10.1073/pnas.1320771110

Wang-Lin, S. X., Olson, R., Beanan, J. M., MacDonald, U., Russo, T. A., and Balthasar, J. P. (2019). Antibody dependent enhancement of *Acinetobacter baumannii* infection in a mouse pneumonia model. *J. Pharmacol. Exp. Ther.* 368, 475–489. doi: 10.1124/jpet.118.253617

Weedall, G. D., and Conway, D. J. (2010). Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends Parasitol.* 26, 363–369. doi: 10.1016/j.pt.2010.04.002

WHO (2019). *World Malaria Report 2019.* Available online at: https://www.who.int/news-room/feature-stories/detail/world-malaria-report-2019.

Winarski, K. L., Tang, J., Klenow, L., Lee, J., Coyle, E. M., Manischewitz, J., et al. (2019). Antibody-dependent enhancement of influenza disease promoted by increase in hemagglutinin stem flexibility and virus fusion kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15194–15199. doi: 10.1073/pnas.1821317116

Woolley, I. J., Wood, E. M., Sramkoski, R. M., Zimmerman, P. A., Miller, J. P., and Kazura, J. W. (2005). Expression of Duffy antigen receptor for chemokines during reticulocyte maturation: using a CD71 flow cytometric technique to identify reticulocytes. *Immunohematology* 21, 15–20.

Wootton, J. C., Feng, X., Ferdig, M. T., Cooper, R. A., Mu, J., Baruch, D. I., et al. (2002). Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418, 320–323. doi: 10.1038/nature00813

World Health Organization (2018). Malaria vaccine: WHO position paper, January 2016 - recommendations. *Vaccine* 36, 3576–3577. doi: 10.1016/j.vaccine.2016.10.047

Wu, J., Cai, B., Sun, W., Huang, R., Liu, X., Lin, M., et al. (2015). Genome-wide analysis of host-*Plasmodium yoelii* interactions reveals regulators of the type i interferon response. *Cell Rep.* 12, 661–672. doi: 10.1016/j.celrep.2015.06.058

Wu, J., Tian, L., Yu, X., Pattaradilokrat, S., Li, J., Wang, M., et al. (2014). Strain-specific innate immune signaling pathways determine malaria parasitemia dynamics and host mortality. *Proc. Natl. Acad. Sci. U.S.A.* 111, E511–E520. doi: 10.1073/pnas.1316467111

Wu, J., Xia, L., Yao, X., Yu, X., Tumas, K. C., Sun, W., et al. (2020). The E3 ubiquitin ligase MARCH1 regulates antimalaria immunity through interferon signaling and T cell activation. *Proc. Natl. Acad. Sci. U.S.A.* 117, 16567–16578. doi: 10.1073/pnas.2004332117

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *PLoS ONE* 6:e26960. doi: 10.1371/journal.pone.0026960

Xia, L., Wu, J., Pattaradilokrat, S., Tumas, K., He, X., Peng, Y. C., et al. (2018). Detection of host pathways universally inhibited after *Plasmodium yoelii* infection for immune intervention. *Sci. Rep.* 8:15280. doi: 10.1038/s41598-018-33599-1

Xu, R., Liu, Y., Fan, R., Liang, R., Yue, L., Liu, S., et al. (2019). Generation and functional characterisation of *Plasmodium yoelii* csp deletion mutants using a microhomology-based CRISPR/Cas9 method. *Int. J. Parasitol.* 49, 705–714. doi: 10.1016/j.ijpara.2019.04.003

Yam, X. Y., Brugat, T., Siau, A., Lawton, J., Wong, D. S., Farah, A., et al. (2016). Characterization of the *Plasmodium* interspersed repeats (PIR) proteins of *Plasmodium chabaudi* indicates functional diversity. *Sci. Rep.* 6:23449. doi: 10.1038/srep23449

Yam, X. Y., and Preiser, P. R. (2017). Host immune evasion strategies of malaria blood stage parasite. *Mol. Biosyst.* 13, 2498–2508. doi: 10.1039/C7MB00502D

Zhang, X., Alexander, N., Leonardi, I., Mason, C., Kirkman, L. A., and Deitsch, K. W. (2019). Rapid antigen diversification through mitotic recombination in the human malaria parasite *Plasmodium falciparum*. *PLoS Biol.* 17:e3000271. doi: 10.1371/journal.pbio.3000271

Zhu, S. J., Hendry, J. A., Almagro-Garcia, J., Pearson, R. D., Amato, R., Miles, A., et al. (2019). The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria. *Elife* 8:e40845. doi: 10.7554/eLife.40845

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.

Copyright © 2020 Su, Zhang and Joy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

cyt b, cytochrome oxidase b gene; ADE, antibody dependent enhancement; AMA1, apical membrane antigen 1; CM, cerebral malaria; CSP, circumsporozoite protein; DBL, Duffy binding-like; CMAH, cytidine monophospho-N-acetylneuraminic acid hydroxylase; *dhfr-ts*, dihydrofolate reductase-thymidylate synthase gene; DTP, diphtheria-tetanus-pertussis; DARC, Duffy antigen/receptor for chemokines; EPCR, endothelial protein C receptor; EBL, erythrocyte binding-like; ECM, experimental cerebral malaria; GPCR, G-protein coupled receptor; GWAS, genome-wide association studies; GPLS, genome-wide patterns of LOD score; GYPA, glycophorin A; HBsAg, hepatitis B virus surface antigen; HTMV, high-titer measle vaccine; HKA, Hudson-Kreitman-Aguade; HLA, human leukocyte antigen; iRBC, infected erythrocyte; *IntRh5,* introgressed Rh5 gene; LGS, linkage group selection; MSP1, merozoite surface protein 1; MSP2, merozoite surface protein 2; MMEJ, microhomology-mediated end joining; MS, microsatellite; MM, mild malaria; OR, olfactory receptor; PfCRT, *P. falciparum* chloroquine resistance transporter; PfDHFR-TS, *P. falciparum* dihydrofolate reductase-thymidylate synthase; PfDHPS, *P. falciparum* dihydropteroate synthase; PfEbl165, *P. falciparum* EBL165 gene; PfEMP1, *P. falciparum* erythrocyte membrane protein 1; N67, *Plasmodium y. nigeriensis* N67; N67C, *Plasmodium y. nigeriensis* N67C; PyEBL, *Plasmodium yoelii* erythrocyte binding-like protein; PAMPs, pathogen-associated molecular patterns; PS, phosphatidylserine; PPIPs, Piroplasm-interacting proteins; *pir*, *Plasmodium* interspersed repeat; PCV, pneumococcal conjugate vaccine; RBC, red blood cell; *rif*, repetitive interspersed; RH5, reticulocyte-binding-like homologous protein 5; RLR, retinoic acid-inducible gene-I-like receptor; SICA, schizont infected cell agglutination; NCM, severe non-cerebral malaria; SNP, single nucleotide polymorphism; *stevor*, subtelomeric variant open reading frame; SPR, surface plasmon resonance; TjD, Tajima's D; CyRPA, cysteine-rich protective antigen; TLR, Toll-like receptors; Ts-eQTL, trans-species expression quantitative trait locus; VE, vaccine efficacy; WHO, World Health Organization.

frontiers in Microbiology

REVIEW
published: 20 November 2020
doi: 10.3389/fmicb.2020.581543

Epidemiological Characteristics and Formation Mechanisms of Multidrug-Resistant Hypervirulent *Klebsiella pneumoniae*

Miran Tang[1†], Xin Kong[2†], Jingchen Hao[2] and Jinbo Liu[2]*

[1] Clinical Laboratory Department, Santai People's Hospital, Mianyang, China, [2] Department of Laboratory Medicine, Affiliated Hospital of Southwest Medical University, Luzhou, China

Multi-drug resistance (MDR) and hypervirulence (hv) were exhibited by different well-separated *Klebsiella pneumoniae* lineages in the past, but their convergence clones—MDR-hypervirulent *K. pneumoniae* (HvKPs)—both highly pathogenic and resistant to most available antibiotics, have increasingly been reported. In light of the clonal lineages and molecular characteristics of the studied MDR-HvKP strains found in the literature since 2014, this review discusses the epidemiology of MDR-HvKPs, in particular summarizing the three general aspects of plasmids-associated mechanisms underlying the formation of MDR-HvKPs clones: MDR-classic *K. pneumoniae* (cKPs) acquiring hv plasmids, hvKPs obtaining MDR plasmids, and the acquisition of hybrid plasmids harboring virulence and resistance determinants. A deeper understanding of epidemiological characteristics and possible formation mechanisms of MDR-HvKPs is greatly needed for the proper surveillance and management of this potential threat.

Keywords: *Klebsiella pneumoniae*, multi-drug resistance, hypervirulent, epidemiology, formation mechanism, plasmid, horizontal gene transfer, mobile genetic elements

OPEN ACCESS

Edited by:
Jianying Gu,
College of Staten Island,
United States

Reviewed by:
Chang-Ro Lee,
Myongji University, South Korea
Andrés Esteban Marcoleta,
University of Chile, Chile

***Correspondence:**
Jinbo Liu
liulab2019@163.com

† 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: *09 July 2020*
Accepted: *14 October 2020*
Published: *20 November 2020*

Citation:
Tang M, Kong X, Hao J and Liu J (2020) Epidemiological Characteristics and Formation Mechanisms of Multidrug-Resistant Hypervirulent Klebsiella pneumoniae. Front. Microbiol. 11:581543. doi: 10.3389/fmicb.2020.581543

INTRODUCTION

Klebsiella pneumoniae is a clinically relevant opportunistic pathogen that causes a wide range of infections. Hypervirulent *K. pneumoniae* (hvKPs) and "classic" *K. pneumoniae* (cKPs) are two different variants of *K. pneumoniae* (Bialek-Davenet et al., 2014). The former are usually hypermucoviscous (HM) and are clinically characterized by their abilities to cause life-threatening invasive community-acquired infections, like entophthalmias and liver abscesses, in a healthy population (Shon et al., 2013). Fortunately, the majority of them have retained susceptibility to multiple commonly used antimicrobial agents except for ampicillin. The latter generally behave as opportunistic avirulent pathogens, causing nosocomial infections in hospitalized patients. Unfortunately, they have a propensity to acquire multiple resistant determinants, such as extended-spectrum β-lactamases (ESBLs) and/or carbapenemase, to be multi-drug resistant (MDR) strains making treatment more difficult (Navon-Venezia et al., 2017).

In the past, virulence and antibiotic-resistance have evolved separately in these two distinct *K. pneumoniae* clonal groups (Bialek-Davenet et al., 2014). However, in the face of antibiotic selection pressure, more and more isolates with combined hypervirulence (hv) and MDR have been

detected and have reached public attention as "superbugs" with the possibility of causing untreatable invasive infections. To better understand such strains, this review focuses on summarizing and discussing the epidemiological characteristics and the possible formation mechanisms of them.

PATHOGEN AND EPIDEMIOLOGY

Over the past years, a "reference/standard" genotypic/phenotypic marker for hvKP has been lacking, and a positive string test indicating HM has been regarded as an important *in vitro* parameter for hvKP identification in some early studies (Yao et al., 2015). Yet, as research progressed, several controversies regarding the association of HM phenotype and hv have been raised. HM is not exclusive to hv, on the contrary, hvKP does not absolutely present HM (Carlos Catalan-Najera et al., 2017). In this review, studies involving hv (well-defined by virulence-related assays) or HM (positive for string test) strains were included.

Klebsiella pneumoniae has experienced the evolution of third-generation cephalosporin-, carbapenem-, and even polymyxin-resistance. Harboring a wide range of β-lactamases results in third-generation cephalosporin-resistance in *K. pneumoniae* isolates. Acquired resistance to carbapenems can be conferred by carbapenemase production, such as KPC, OXA, and MBLs which include NDM, VIM, IMP, as well as SIM (Tzouvelekis et al., 2012). Then, the prevalence of carbapenem-resistant bacteria has led to the use of polymyxins as a last-therapy option to treat associated infections, which has resulted in the emergence of polymyxins-resistant *K. pneumoniae*. This review includes literature about any kind of MDR phenotypes hvKPs.

We performed an exhaustive search of PubMed, MEDLINE, Web of Science, EMBASE, CNKI, and Wanfang database for English-language literature published before January, 2020, with the following search strategy ("hypervirulence"[All Fields] OR "hypervirulent"[All Fields] OR "hypermucoviscous"[All Fields]) AND ("*Klebsiella pneumoniae*"[MeSH Terms] OR "*Klebsiella pneumoniae*"[All Fields]) AND ("resistance"[All Fields] OR "resistant"[All Fields]), then summarized and classified these papers according to the country or region, STs, capsule types, plasmid replicon types, resistance loci, and formation mechanisms in **Table 1**, which we will refer to frequently throughout this section.

From **Table 1**, we can see that MDR-HvKPs have mainly been detected since 2014 and have become research hotspot. All studies since 2014 involving both MDR and hv *K. pneumoniae* isolates account for a total of 47. China (including Taiwan) accounts for 33, and the remaining 13 are from other Asian countries (three from India, one from Japan, and one from Iran), Europe (two from the United Kingdom, one from France, one from Italy, and one from Norway), North America (one from United States), and South America (one from Brazil and another from Argentina). In our opinion, such geographical distribution is attributed to the prevalence characteristics of MDR and hv-KP strains around the world. For example, the high prevalence of both hvKPs and MDR-KPs in China and the significant proportions of incidence of MDR-KPs in other Asian countries (Lee et al., 2016, 2017) may contribute to the majority of reports about MDR-HvKPs coming from these regions.

By comparing the allelic sequences of seven housekeeping genes, multi-locus sequence typing (MLST) can structure *K. pneumoniae* populations into lineages, which are typically referenced by their sequence types (STs; e.g., ST11). The common MDR-KP strains are strongly linked to particular clonal complexes (CCs), like CC258, comprising ST258, ST11, ST512, ST340, ST437, etc. (Schweizer et al., 2019), CC15 and CC14, while hvKPs mainly belong to ST23 for the K1 capsular serotype and to ST86, ST65, and ST25 for K2 (Bialek-Davenet et al., 2014). Hence, the genetic backgrounds of the isolates, i.e., the strains stem from whether resistant or virulent lineage, can be determined by STs. MDR-HvKPs showed various STs in the literature (**Table 1**). Among them, the most prevalent ST was the KPC-producing CRKP highly related ST11, followed by the dominant hvKP lineage ST23. This was in accordance with the fact that ST11 and ST23 are predominant among MDR-KP and hvKP strains, respectively (Bialek-Davenet et al., 2014). In addition, STs showed different region distributions. In China, ST23 accounted for about 69.57% (16/23) in all hvKPs and ST11 took on almost the same high proportion 77.27% (51/66) in all MDR-KPs, respectively, which was consistent with the fact that both KPC-KP and hvKP are prevalent in China (Lee et al., 2016, 2017). While in other countries, except for India and Brazil, ST11 was almost undetected; instead, STs representing a virulent clone, such as ST23, ST25, and ST86, were more easily detected, which was in line with the fact that hvKP has increasingly prevailed in Europe and the Americas in recent years (Cubero et al., 2016; Lee et al., 2017).

POSSIBLE FORMATION MECHANISMS

Bacterial phenotypic changes including resistance and virulence acquisition are mainly driven by horizontal gene transfer (HGT) (in addition to chromosomal mutations) mediated by mobile genetic elements (MGEs), such as plasmids, insertion sequences (IS), transposons (Tn), integrons (In) and integrative conjugative elements (ICEs), driving dissemination and co-selection of virulence and resistance genes through genomic rearrangement during their replication or recombination process, mostly in Gram-negative pathogens including *K. pneumoniae* (Frost et al., 2005). Compared with transformation and transduction, thought to be secondary effects of other biological processes, conjugation plays a most impactful role in HGT due to the transfer of plasmids and ICEs *via* direct cell-to-cell contact (Pinilla-Redondo et al., 2018). Additionally, in most instances, hv or MDR phenotype selection genes transferred by Tn, In, and IS were mostly accumulated on plasmids followed by plasmid inter bacterial transfer. Consequently, as the most pivotal vectors for conjugation and genes recruitment, plasmids are arguably the most indispensable and essential elements throughout the entire convergence process of hv and MDR genes coming from strains of distinct backgrounds. Accordingly, in this section, we dissect the available mechanism studies apart from case reports

TABLE 1 | Lists of papers referenced.

Country	Sequence type(s)	Capsule type(s)	Resistance mechanism(s)	Plasmid replicon type(s)	Formation mechanisms*	References
China	ST29 ($n = 1$)	K54 ($n = 1$)	bla_{NDM-5} ($n = 1$)	p^{vir}IncHI1/IncFIB p^{res}IncX3	oo	Yuan et al., 2019
	ST1764 ($n = 7$) ST11 ($n = 6$)	K64 ($n = 7$) K47 ($n = 4$), K64 ($n = 2$)	bla_{NDM-1} ($n = 7$) bla_{KPC-2} ($n = 6$)	NA	NA	Liu Z. et al., 2019
	ST86 ($n = 1$)	K2 ($n = 1$)	bla_{NDM-1}, bla_{KPC-2} co-carrying ($n = 1$)	p^{vir}IncHI1/IncFIB p^{res}IncFII(K) and p^{res}IncN	oo	Liu Y. et al., 2019
	ST11 ($n = 12$) ST23 ($n = 1$) ST660 ($n = 1$) ST1660 ($n = 1$)	K64 ($n = 7$), K47 ($n = 5$) K1 ($n = 1$) K16 ($n = 1$) K1 ($n = 1$)	bla_{KPC-2} ($n = 15$)	NA	NA	Xu et al., 2019
	ST65 ($n = 1$)	K2 ($n = 1$)	$bla_{CTX-M-3}$, $bla_{CTX-M-14}$ coharboring	NA	NA	Fu et al., 2019
	ST23 ($n = 1$)	K1 ($n = 1$)	bla_{VIM-1} ($n = 1$)	p^{vir}IncHI1B/IncFIBk p^{res}IncA and p^{res}IncFII	oo	Dong et al., 2019
	ST23 ($n = 1$)	K1 ($n = 1$)	bla_{NDM-1} ($n = 1$)	NA	oo	Liu and Su, 2019
	ST25 ($n = 16$) ST11 ($n = 3$) ST375 ($n = 1$)	K2 ($n = 16$) Non-typeable ($n = 3$) K2 ($n = 1$)	bla_{KPC-2} ($n = 10$), bla_{NDM-1} ($n = 1$), $ESBL_S$ ($n = 5$) bla_{KPC-2} ($n = 3$) $ESBL_S$ ($n = 1$)	NA	NA	Li et al., 2019
	ST15 ($n = 7$)	KL112 ($n = 7$)	$bla_{OXA-232}$ ($n = 7$)	$p^{res-OXA}$ColKP3-type $p^{res-CTX}$ IncFII $p^{res-MDR}$IncFIB p^{vir}IncHI1B/IncFIB	oo	Shu et al., 2019
	ST23 ($n = 1$)	K1 ($n = 1$)	$bla_{CTX-M-24}$ ($n = 1$)	$p^{vir-CTXM}$ IncHI1B/IncFIB	o	Shen et al., 2019
	ST23 ($n = 2$) ST412 ($n = 1$) ST660 ($n = 1$) ST700 ($n = 1$)	K1 ($n = 2$) K57 ($n = 1$) K16 ($n = 1$) K1 ($n = 1$)	Undefined ($n = 2$) PhoQD150G ($n = 1$) PhoQD150G ($n = 1$) mcr-1 and PhoQD150G ($n = 1$)	NA	NA	Lu et al., 2018
	ST23 ($n = 1$)	K1 ($n = 1$)	bla_{DHA-1} ($n = 1$)	$p^{res-DHA}$IncHI5 p^{vir}IncHI1B/IncFIB	oo	Xie et al., 2018
	ST2922 ($n = 1$)	K1 ($n = 1$)	bla_{DHA} and $bla_{CTX-M-14}$ ($n = 1$)	p^{res}IncR p^{vir}IncFIB/IncHI1B	NA	Xu et al., 2018
	ST36 ($n = 1$)	K62 ($n = 1$)	bla_{KPC-2} ($n = 1$)	$p^{ress-KPC}$ IncFII p^{vir} IncHI1/IncFIB	oo	Feng et al., 2018
	ST11 ($n = 1$)	K47 ($n = 1$)	bla_{KPC-2} ($n = 1$)	$p^{ress-KPC}$IncR,IncFII, IncNp$^{res-MDR}$ IncA/C2 p^{vir} IncHI1B/IncFIB	oo	Huang et al., 2018
	Unknown ($n = 18$)	K1 ($n = 18$)	$ESBL_S$ and overexpression of efflux pumps ($n = 18$)	NA	NA	Lin et al., 2018
	ST86 ($n = 1$)	K2 ($n = 1$)	bla_{NDM-1}- and bla_{KPC-2}- Coproducing ($n = 1$)	NA	NA	Wei et al., 2018
	ST11 ($n = 3$)	Unknown ($n = 3$)	bla_{KPC-2} or bla_{NDM-1} ($n = 3$)	NA	NA	Wong et al., 2018
	ST11 ($n = 1$)	Unknown ($n = 1$)	tet(A) variant and bla_{KPC-2} ($n = 1$)	NA	o	Gu et al., 2018
	ST11 ($n = 3$)	K47 ($n = 3$)	bla_{KPC-2} ($n = 3$)	NA	oo	Gu et al., 2017 Dong et al., 2018b
	ST11 ($n = 4$)	Unknown ($n = 4$)	tet(A) variant and bla_{KPC-2} ($n = 1$)	NA	oo	Yao et al., 2018

(Continued)

TABLE 1 | Continued

Country	Sequence type(s)	Capsule type(s)	Resistance mechanism(s)	Plasmid replicon type(s)	Formation mechanisms*	References
	ST11 ($n = 16$) ST268 ($n = 2$) ST65 ($n = 1$) ST692 ($n = 1$) ST595 ($n = 1$)	K20($n = 5$),non-T($n = 11$) K20($n = 1$),non-T($n = 1$)K2 ($n = 1$) non-T ($n = 1$) non-T ($n = 1$)	bla_{KPC-2} and *bla*SHV-11 ($n = 16$) bla_{KPC-2} and *bla*SHV-11 ($n = 2$) bla_{KPC-2} and *bla*SHV-11 ($n = 1$) bla_{KPC-2} and *bla*SHV-11 ($n = 1$) bla_{KPC-2} ($n = 1$)	NA	NA	Zhan et al., 2017
	ST23 ($n = 2$) ST268 ($n = 3$) ST65 ($n = 1$) ST17 ($n = 1$) ST420 ($n = 1$) ST367 ($n = 1$) ST1658 ($n = 1$) ST35 ($n = 1$)	K1 ($n = 2$) K1 ($n = 1$),K20 ($n = 2$) K2 ($n = 1$) Non-typeable ($n = 1$) K20 ($n = 1$) K1 ($n = 1$) K2 ($n = 1$) Non-typeable ($n = 1$)	SHV-75, CTXM-55, SHV-11, TEM-1, CTX-M-like, SHV-148, CTX-M-14, TEM-53 ($n = 11$)	NA	NA	Zhang Y.W. et al., 2016
	ST86 ($n = 7$) ST37 ($n = 6$) ST23 ($n = 5$)	Unknown	$ESBL_S$ ($n = 18$)	NA	NA	Zhang J. et al., 2016
	ST661 ($n = 1$)	K1 ($n = 1$)	mcr-1 ($n = 1$)	NA	NA	Gu et al., 2016
	ST14 ($n = 1$)	K2 ($n = 1$)	bla_{NDM-5} ($n = 1$)	NA	NA	Liu et al., 2016
	ST11 ($n = 1$)	K1 ($n = 1$)	bla_{KPC-2} ($n = 1$)	$p^{virs\text{-}KPC}$IncFIIk	∘	Wei et al., 2016
	ST23 ($n = 1$)	K1 ($n = 1$)	bla_{KPC-2} ($n = 1$)	$p^{virs\text{-}KPC}$IncHI1B/ IncFIB	∘	Zhang R. et al., 2015
	ST23 ($n = 1$) ST1797 ($n = 3$)	K1 ($n = 1$) K1 ($n = 3$)	bla_{KPC-2} ($n = 1$) bla_{KPC-2} ($n = 3$)	NA NA	∘∘ ∘	Dong et al., 2018a
	ST65 ($n = 1$)	K2 ($n = 1$)	SHV-11,TEM-53-producing ompK35,36 decreased ($n = 1$)	NA	NA	Zhang Y.W. et al., 2015
	ST25 ($n = 2$) ST65 ($n = 5$) ST11 ($n = 1$)	K2 ($n = 2$) K2 ($n = 5$) Non-typeable ($n = 1$)	bla_{KPC-2} ($n = 6$)	NA	NA	Yao et al., 2015
	Unknown ($n = 5$)	K1 and K2 ($n = 5$)	$ESBL_S$ ($n = 5$)	NA	NA	Li et al., 2014
United Kingdom	ST101 ($n = 3$) ST383 ($n = 3$) ST147 ($n = 4$) ST15 ($n = 2$) ST48 ($n = 1$)	Unknown	*bla*OXA-48 and MDR ($n = 1$), MDR ($n = 2$) *bla*OXA-48 ($n = 1$), bla_{NDM-5} ($n = 2$), MDR ($n = 1$) bla_{NDM-1} and MDR ($n = 2$), MDR ($n = 2$) *bla*NDM and MDR ($n = 1$), MDR ($n = 1$) bla_{NDM-5} and MDR ($n = 1$)	$p^{vir\text{-}MDR\text{-}OXA\text{-}48}$ and (or) -NDM IncFII(K)/IncFIB(K) $p^{res\text{-}OXA\text{-}48}$IncL/M $p^{res\text{-}NDM\text{-}1}$IncFIB(pQil) $p^{vir\text{-}NDM\text{-}5}$ IncFIB(Mar)	∘	Turton et al., 2017, 2019
Argentina	ST25 ($n = 1$)	K2 ($n = 1$)	bla_{KPC-2} ($n = 1$)	NA	NA	Cejas et al., 2019
Japan	ST23 ($n = 1$)	K1 ($n = 1$)	*bla*IMP-6 ($n = 1$)qq	$p^{res\text{-}IMP\text{-}6}$IncN	∘∘	Harada et al., 2019
Norway	ST15 ($n = 2$)	K24 ($n = 2$)	$ESBL_S$ ($n = 2$)	$p^{vir\text{-}CTXM}$ $IncFIB_{K_1}$/$IncFII_{K_1}$	∘	Lam et al., 2019
Iran	ST23 ($n = 5$)	K1 ($n = 5$)	*bla*VIM-2 ($n = 5$)	$p^{res\text{-}VIM-2}$IncN	∘∘	Tabrizi et al., 2018
United Kingdom	ST23 ($n = 1$)	K1 ($n = 1$)	bla_{NDM-1} ($n = 1$)	NA	NA	Roulston et al., 2018
Italy	ST512 ($n = 1$)	Unknown ($n = 1$)	*bla*KPC-3 ($n = 1$)	NA	NA	Arena et al., 2017
France	ST86 ($n = 1$)	K2 ($n = 1$)	$bla_{CTX\text{-}M\text{-}3}$ ($n = 1$)	p^{res}IncL/M	NA	Surgers et al., 2016
India	ST2318 ($n = 1$)	Non-typeable ($n = 1$)	$ESBL_S$ ($n = 1$)	NA	NA	Shankar et al., 2016b

(Continued)

TABLE 1 | Continued

Country	Sequence type(s)	Capsule type(s)	Resistance mechanism(s)	Plasmid replicon type(s)	Formation mechanisms*	References
India	ST11 ($n = 1$) ST43 ($n = 1$) ST231 ($n = 1$)	Unknown	$bla_{OXA\text{-}232}$, $bla_{OXA\text{-}181}$, $bla_{OXA\text{-}1}$, $bla_{NDM\text{-}1}$	IncFIA IncFIB IncFII IncHI1B Col	NA	Shankar et al., 2016a
India	ST14 ($n = 1$)	Unknown ($n = 1$)	Mutation in OmpK36 ($n = 1$)	NA	NA	Rafiq et al., 2016
United States	ST23 ($n = 1$)	Unknown ($n = 1$)	$bla_{KPC\text{-}2}$ ($n = 1$)	$p^{res\text{-}KPC}$IncFIA	NA	Cejas et al., 2014
Brazil	ST11 ($n = 7$)	Unknown ($n = 7$)	$bla_{KPC\text{-}2}$ qnrS1 *bla*CTX-M-2 ($n = 7$)	$p^{ress\text{-}KPC}$IncFIIk	NA	Andrade et al., 2014

**Single ∘ represents acquiring hybrid plasmid by KPs, double ∘∘ represents acquiring additional plasmid(s), p^{vir} by MDR KPs or p^{res} by hvKPs, respectively.*

(**Table 1**) to extract and further expand to three general aspects of plasmids-associated mechanisms underlying the convergence of hv and MDR phenotypes.

MDR-cKPs Acquire Hypervirulence Plasmids

Virulence plasmid acquisition is an important mechanism for the increased virulence of MDR-cKPs. The best characterized virulence plasmids are the 224 kbp plasmid pK2044 from K1, ST23 strain NTUH-K2044; the 219 kbp plasmid pLVPK from K2, ST86 strain CG43; and the 121 kbp plasmid Kp52.145pII from K2, ST66 strain Kp52.145, on which the virulence-associated loci and genes were highly conservatively organized (Lam et al., 2018). There are some convincing and rational explanations to the comparative rarity of hv plasmids in avirulent MDR-KP clones compared to in hvKPs. First, dominant hv plasmids hosting KP ST23 accounts for only $\leq$ 2% of clinical *K. pneumoniae* isolates in the global range, except for in the Asia-Pacific rim, despite circulating among humans with hv plasmid maintenance for more than 100 years, far longer than the most well-known MDR clones (Wyres et al., 2020), suggesting that the hv plasmids rarely move horizontally and are highly restricted to hvKP clones, i.e., lineage specificity of plasmid distribution. Second, large numbers of plasmids, like hv plasmids in *K. pneumoniae*, devoid of genes permitting their transfer by conjugation and relying on the transfer function of other plasmids to enable their transfer, are mobilizable but not self-transmissible (Smillie et al., 2010), which radically limits the frequency of the horizontal transfer of hv plasmids. Third, as large low-copy-number plasmids, virulence plasmids contain specific replication and maintenance systems to ensure their transmission to daughter cells of specific genetic backgrounds (Million-Weaver and Camps, 2014), which is reminiscent of the fact that they might often impose fitness costs on other unsuitable host. Despite those restrictive factors, in fact, MDR-cKPs acquiring virulence plasmids have recently been reported. Yao et al. (2018) screened four ST11 CR-HvKP strains from clinical patients in Henan province, China, each of which carried both a KPC-2-encoding and a virulence plasmid. Further sequencing of the virulence plasmid showed high homology to pLVPK. Resembling that, Gu et al. (2018) reported the emergence of a tigecycline- and carbapenem-co-resistant ST11 hvKP isolate from a patient's gut in Zhejiang, China. One of its three plasmids shared high homology with pLVPK and another co-carried $bla_{KPC\text{-}2}$ and *tet*(A). In the Lancet, Gu et al. reported a fatal outbreak of ST11 CR-HvKP strains in a Chinese hospital. In addition to their intrinsic conjugative MDR plasmids carrying $bla_{KPC\text{-}2}$ genes, the five representative causative strains acquired an additional virulence plasmid that aligned well to most parts of pLVPK (Gu et al., 2017). Subsequently, Dong et al. made the whole genome sequences (WGS) of three ST11 CR-HvKP isolates surveyed in the previous Gu et al.'s study and five plasmids harbored by each of them. The presence of homologous regions between the virulence plasmid and $bla_{KPC\text{-}2}$-bearing conjugative MDR plasmid suggested that their co-integrated transfer might mediate the transmission of the non-conjugative virulence plasmid from hvKP to ST11 CRKP (Dong et al., 2018b). The main explanation for these facts is that the virulence plasmids are not self-transmissible but often mobilized to access new hosts with the help of other ICEs or conjugative resistance plasmids encoding the conjugation transfer complex in the same host cell (Ramsay and Firth, 2017). Of note, the parental hvKPs, typically susceptible to antimicrobial agents, generally do not originally carry the conjugative resistance plasmid, so in this scenario, the event that hv plasmids are synergistically mobilized by conjugative resistance plasmids in the same host cell to new strains happens after the conjugative resistance plasmids are transferred into hvKPs first, as will be discussed in the next section.

Hv (HM) KPs Obtain Multidrug-Resistance Plasmids

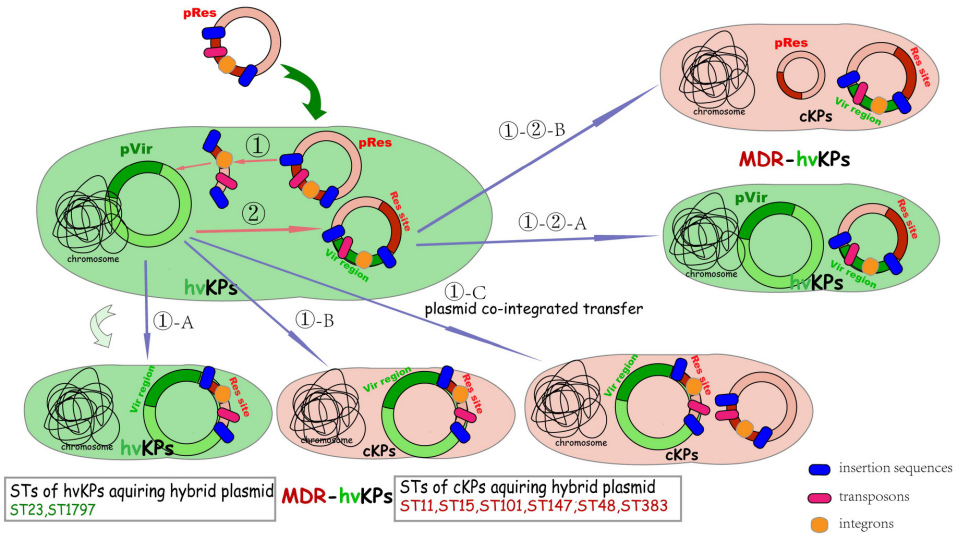
From the comparatively lower genome's diversity in hvKPs than MDR-KPs, Wyres et al. (2019) inferred that hv clones are less likely to acquire resistance genes than MDR clones are to acquire virulence genes. Nevertheless, evidence seemly shows the opposite. We found that out of all the documents reviewed, the total number of Hv (HM) KP isolates obtaining MDR plasmids was much more than that of MDR-cKPs acquiring hv plasmids, 111 *vs.* 87, respectively. (Not absolutely, but it seems to be a doubt). In Zhang R. et al.'s (2015) study, both carbapenems-susceptible HvKP and CR-HvKP strains harbored an ∼200-kb virulence plasmid but the latter had acquired two additional plasmids with $bla_{KPC\text{-}2}$ gene located on a transferable

plasmid. Tabrizi et al. described the emergence of VIM-2-encoding K1 ST23 CR-HvKP in an outbreak in Iran. Plasmid analysis revealed a class 1 In carrying bla_{VIM-2} located on an ∼45-kb IncN conjugative plasmid (Tabrizi et al., 2018). As reported by Feng et al. (2018), a bla_{KPC-2}-mediated carbapenem-resistant ST36 hvKP clinical isolate had two plasmids, one IncHI1/IncFIB plasmid highly similar to the known pLVPK, another IncFII plasmid carrying bla_{KPC-2} and proved self-transmissible. Similarly, the first IMP-producing K1 ST23 CR-HvKP in Japan carried a pLVPK-like plasmid and an IncN plasmid harboring class 1 In-mediated bla_{IMP-6}. According to the fact that bla_{IMP-6} was, while *rmpA* was not, detected in the transconjugant, the authors inferred that the bla_{IMP-6}-carrying plasmid was conjugative, but the plasmid carrying virulence gene was not, and they were two individual plasmids (Harada et al., 2019). Liu et al. presented an NDM-1 and KPC-2 co-producing K2 ST86 CR-HvKP strain with four plasmids in China. Apart from an IncHI1/IncFIB virulence plasmid identical to pLVPK, the strain additionally acquired two carbapenemase-producing plasmids including bla_{NDM-1}-carrying IncN plasmid and IncFIIK plasmid which carried bla_{KPC-2} and an array of other resistance elements (Liu Y. et al., 2019). Xie et al. delineated a bla_{DHA-1}-carrying IncHI5 plasmid which had a 26-kb accessory region where the bla_{DHA-1} gene was located upstream of ISCR1 isolated from a K1 ST23 MDR-HvKP strain. Apart from this MDR plasmid, this strain carried another two plasmids, including its virulence plasmid (Xie et al., 2018). In Dong et al. (2019) study, a VIM-1-producing K1 ST23 CR-HvKP strain harbored three plasmids; a virulence plasmid highly homologous to that recovered from other ST23 hvKPs; and a bla_{VIM-1}-bearing plasmid possessing a unique resistance island structure presumably generated by multiple gene mobilization events. Yuan et al. (2019) showed that apart from an IncHI1/IncFIB pLVPK-like plasmid, a bla_{NDM-5}-carrying K54 ST29 CR-HvKP isolated from Sichuan, China, harbored a bla_{NDM-5}-carrying IncX3 self-transmissible plasmid. Recent WGS work of an NDM-1-producing K1 ST23 CR-HvKP in China, by Liu and Su (2019), showed that in addition to a pLVPK-like virulence plasmid, it had a conjugative resistance plasmid carrying a bla_{NDM-1} and another six types of resistance genes surrounded by ISs. All of these studies were typical examples in which hvKP clones additionally acquired resistance plasmids all proved conjugative. On the theoretical basis that most large resistance plasmids encode their own transfer and are conjugative, we could further speculate from these papers that the surveyed phenotypically convergent strains originally carried a virulence plasmid and thereafter acquired extra resistance plasmids, which are readily transmitted by horizontal transfer between different lineages and species. Nevertheless, if we only consider the strong transfer of resistance plasmids, we cannot explain why the composite strains are still far less than MDR-KP strains. Therefore, we guess that the success of the MDR and hv convergence results from interaction of various positive and negative factors. Clear examples of the positive factors include the following: (i) most resistance plasmids encode all functions needed for their horizontal transfer, such as DNA replication and copy number control functions, mating pair formation genes, and an origin of transfer (oriT) (Pinilla-Redondo et al., 2018), which facilitated their transfer into other strains like hvKPs; (ii) acquisition of antibiotic resistance genes to become MDR-HvKPs promotes adaptive evolution of hvKP clones in an antibiotic environment; (iii) most acquisition of antibiotic resistance will reduce the virulence and fitness of the strain, especially in the absence of antibiotic selection (Durao et al., 2018); while in the era of antibiotics, antibiotic selection pressure promotes plasmids persistence once resistant mutants form; (iv) conjugative resistance plasmids are usually large and have a low-copy-number, which to some extent circumvent use of host material, and as such, has less effect on host fitness than high-copy-number ones; (v) host genetic background might be a pivotal determinant of plasmid fitness. Strain-dependent compensation to the cost of resistance acquisition might occur in hv strains. Intriguingly, not as many strains of concurrent hv and MDR phenotypes, as imagined in the context of so many positive advantages, implied the essential role of negative factors: (i) considerably lower genome diversity and plasticity than that of MDR-KP is a hint that there might be some sort of barrier for hvKP to uptake and/or integrate DNA fragments; (ii) hv strains typically sensitive to antibiotic are selectively killed by antimicrobial agents before they acquire resistance plasmids. Therefore, antibiotic usage creates conditions for the growth and development of MDR-cKP populations rather than producing a noticeable number of MDR-HvKPs. Altogether, we provide explanations for the complex interactions between resistance plasmids and hvKPs, which shed light on the mechanisms of Hv (HM) KPs obtaining and maintaining MDR plasmids. Notwithstanding Wyres et al.'s (2019) opinion posed at the beginning of this section, that hvKPs are unable or difficult to obtain resistance plasmids, it is possible to state that the genomic traits of hvKPs just weaken the strong transmission kinetics of MDR factors into themselves than into cKPs, producing relatively fewer MDR-hvKPs than MDR-cKPs. Under the interaction of many factors, it is still very easy for hvKPs to obtain resistance plasmids.

Virulence-Resistance Hybrid Plasmids

Small MGEs, such as Tn, IS, or In, that "hitchhike" on the plasmids, especially resistance plasmids (Pinilla-Redondo et al., 2018), contribute to the capture and dissemination of MDR and/or hv genes whose co-existence on the same plasmid constitutes the perfect mosaic structure-hybrid plasmid. The transfer of the hybrid plasmids into either hvKP or cKP strains form MDR-HvKPs. The possible evolution pathway of MDR-HvKPs mediated by the hybrid plasmid is shown in model diagram **Figure 1**.

Some clear examples have been demonstrated. Zhang et al. reported the emergence of five K1 CR-HvKP strains causing fatal infections in hospital patients in Zhejiang Province, China. The K1 ST23 CR-HvKP70-2 harbored an ∼200-Kp plasmid on which bla_{KPC} and *rmpA* were located, and this plasmid was not transferred to *Escherichia coli*. Similarly, bla_{KPC-2} was detected on two virulence genes-harboring plasmids which were not transferable to *E. coli* in other three genetically related K1 ST1797 isolates (Zhang R. et al., 2015). The non-conjugativity



Model diagram of the possible evolution pathway of MDR-hvKPs mediated by hybrid plasmid.

FIGURE 1 | Model diagram of the possible evolution pathway of MDR-hvKPs mediated by a hybrid plasmid. ① MDR plasmids first transfer into hvKPs, then antimicrobial resistance genes are integrated or transposed into hv plasmid harbored by hvKPs, resulting in the formation of hybrid plasmids with most hv genes-bearing regions and the MDR-hvKPs of hv-associated STs **(①-A)**. Alternatively, if the genes encoding the self-transfer conjugative system are integrated into the virulence plasmid, together with the resistance determinants, the hybrid plasmids will be conferred self-transmission and conjugativity transfer into any bacterial host including cKPs to become MDR-HvKPs **(①-B)**. In addition, with the help of other conjugative plasmids this hybrid plasmid can be possibly transferred into other cKPs to form MDR-hvKPs of MDR- or cKP-linked STs **(①-C)**. ② If the In, Tn, and (or) Is further carry hv genes from the hv plasmid into other resistance plasmids, hybrid plasmids with most sites of resistance plasmid characteristics are formed. They can be transferred into either hvKPs **(①-②-A)** or cKPs **(①-②-B)** to form MDR-hvKPs via their conjugal transfer system.

implied that these hv-MDR plasmids were very likely originally harbored by the hvKPs and carried the hv plasmid backbone, then formed mosaic structures *via* MGEs-mediated integration with the bla_{KPC-2}-bearing DNA fragment which comes from other conjugative bla_{KPC-2}-carrying resistance plasmid(s) transferred into the hv host before (**Figure 1**①-A). This hypothesis was subsequently proved in Dong et al.'s study where a hybrid plasmid recovered from CR-HvKP strain KP70-2 was found to be almost structurally identical to numerous known hv conferring plasmids harbored by other hvKP strains, except for an extra MDR-encoding region flanked by two copies of IS26 in the same orientation and MGEs-mediated resistance genes *dfrA14* and bla_{KPC-2}. The authors concluded that multiple IS elements were responsible for the integration of the MDR region into the virulence plasmid (Dong et al., 2018a). Similarly, the complete genome of an ESBL-producing K1 ST23 MDR-HvKP showed that the strain obtained a rare plasmid harboring virulence and $bla_{CTX-M-24}$ genes. Furthermore, a full-plasmid BLAST comparative analysis illustrated that this plasmid exhibited high similarity with three IncHI1B/IncFIB virulent plasmids retrieved from the GenBank, except for a unique $bla_{CTX-M-24}$-harboring region. Further exploration proved that following the IS-mediated $bla_{CTX-M-24}$ gene insertion, into the conserved virulence plasmid backbone region, the hybrid plasmid formed (Shen et al., 2019). In conclusion, these studies suggested that hv strains are capable of acquiring MDR determinants through the integration of the MDR region mediated by MGEs, like Tn, In, and seemly dominant IS, into its intrinsic virulence plasmid. The integration of additional resistance elements into the virulence plasmids of hvKPs constitutes perfect mosaic plasmids possessing dual characteristics of conserved virulent regions and newly acquired MDR-encoding sites, but maybe not conjugative (**Figure 1**①-A).

However, if the genes encoding a self-transfer conjugative system are integrated into the virulence plasmid together with the resistance determinants, the hybrid plasmids will be conferred self-transmission and conjugativity. Alternatively, with the help of other conjugative plasmids, this hybrid plasmid could possibly be mobilized to transfer. Hence, if the fitness cost brought into the host bacteria is not considered, they enable one-time simultaneous transfer of resistance and virulence genes into any type of *K. pneumoniae* clones, including cKPs, to facilitate emergence of MDR-HvKPs (**Figures 1**①-B,①-C). Indeed, a growing body of evidence supports the hypothesis. Huang et al. (2018) identified a hybrid virulent plasmid which comprised both parts of the pLVPK and an IncHI2-type resistance plasmid in a KPC-2-producing K47 ST11 MDR-HvKP strain in Taiwan. According to the study of Lam et al. (2019), both two ESBL-producing K24 ST15 *K. pneumoniae* isolates carried large hv-MDR mosaic plasmids which include sequences typical of $IncFIB_K$ virulence plasmids, such as pK2044, fused with regions of homology with typical IncFIIK conjugative MDR plasmids. Similarly, Turton et al. (2019) also described hybrid plasmids containing

both resistance and virulence clusters in 12 CR-HvKP isolates belonging to ST15/48/101/147/383. Yet, considering the special characteristics of classic hv plasmid, such as its notable lineage specificity, narrow-host-range, and low-copy-number traits, it might inhibit a second virulent plasmid transfer into the same strain, which thus might cause hvKPs to not acquire and maintain other hybrid plasmids with most hv genes-bearing regions. These surveyed CR-HvKP strains of hvKP-linked STs generally carrying a single hv or MDR-hv hybrid plasmid can support this hypothesis.

Besides, if the In, Tn, and/or Is further carry the virulent sites/genes from the hv plasmid into other resistance plasmids, hybrid plasmids with most sites of the resistance plasmid traits are formed. They can be transferred into either hvKPs (**Figure 1**①-②-A) or cKPs (**Figure 1**①-②-B) to form MDR-hvKPs mediated by their conjugal transfer system.

From a biological point of view, the highly mosaic nature of antimicrobial resistance and virulence determinants converging within a single vector, the purported co-selection, efficiently facilitates the evolution in two directions, which is a shortcut to an evolutionary success for a proficient bacteria, since selection by relevant antibiotics will also select for virulence traits (Turton et al., 2019). It should be noted that, due to the integration of multiple plasmids elements, including self-transfer conjugative system as well as expanded replicons number and host ranges, hybrid plasmids with MDR and hv biphenotypes can widely spread and infect many types of bacterial hosts and eventually become notorious environmental contaminators (Xie et al., 2020).

CONCLUSION

In the context of what has already been reported by others regarding the MDR and hv convergence in *K. pneumoniae*, epidemiological characteristics and formation mechanisms of MDR-HvKPs researched in these papers have been discussed and elaborated in the current review.

Collectively, an epidemiology analysis enhances our understanding that the genetic background and geographical distribution characteristics of MDR-HvKP are highly consistent with the epidemic characteristics of hvKP and MDR strains and the monitoring and control of both will help prevent the occurrence of superbugs. The formation mechanism analysis has brought to light that hvKP and MDR strains could evolve into MDR-HvKPs through acquiring MDR, hv plasmids, or resistance/virulence hybrid plasmids. The formation mechanisms of biphenotypic composite bacteria are different in distinct genetic background *K. pneumoniae* clones, but the horizontal transfer of plasmids plays a decisive role. In addition, the adaptive evolution ability of the strain promotes the formation, persistence, and transmission of MDR-HvKP.

In the current review, we identified the important role of plasmid-mediated HGT endowing convergence of hv and MDR in *K. pneumoniae*. A plasmid-centered outlook, as opposed to the traditional host-centric view, should be taken into consideration by clinicians to turn the focal point from the specific host strains to resistance and virulence plasmids (and other MGEs) and to adopt surveillance strategies to track, hinder, or minimize the horizontal dissemination of them. Notably, avoiding drug abuse after the composite strains form may reduce the persistence of plasmids and extreme strains.

AUTHOR CONTRIBUTIONS

MRT and XK conceived, designed, and wrote the manuscript. JCH performed paper search and graphics visualization. JBL performed writing – reviewing and editing, supervision, and project administration. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from (1) Construction of clinical medicine strain resource base and application platform, Sichuan Science and Technology Infrastructure Platform Project (2018TJPT0011) From Science and Technology Department of Sichuan Province, (2) The Study on risk factors and predictive model of high-virulence *Klebsiella pneumoniae* hepatic abscess, Sichuan International Science and Technology Innovation Cooperation/Hong Kong, Macao and Taiwan Science and Technology Innovation Cooperation Project (2019YFH0021) From Science and Technology Department of Sichuan Province, and (3) Proteomics study of carbapenem resistant *Acinetobacter baumannii* (14035) from Luzhou Medical College. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

Andrade, L. N., Vitali, L., Gaspar, G. G., Bellissimo-Rodrigues, F., Martinez, R., and Darini, A. L. (2014). Expansion and evolution of a virulent, extensively drug-resistant (polymyxin B-resistant), QnrS1-, CTX-M- 2-, and KPC-2-producing *Klebsiella pneumoniae* ST11 international high-risk clone. *J. Clin. Microbiol.* 52, 2530–2535. doi: 10.1128/jcm.00088-14

Arena, F., Henrici, De Angelis, L., D'Andrea, M. M., Cannatelli, A., Fossati, L., et al. (2017). Infections caused by carbapenem-resistant *Klebsiella pneumoniae* with hypermucoviscous phenotype: A case report and literature review. *Virulence* 8, 1900–1908. doi: 10.1080/21505594.2017.1286439

Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A.-S., et al. (2014). Genomic Definition of Hypervirulent and Multidrug-Resistant *Klebsiella pneumoniae* Clonal Groups. *Emerg. Infect. Dis.* 20, 1812–1820. doi: 10.3201/eid2011.140206

Carlos Catalan-Najera, J., Garza-Ramos, U., and Barrios-Camacho, H. (2017). Hypervirulence and hypermucoviscosity: Two different but complementary Klebsiella spp. phenotypes? *Virulence* 8, 1111–1123. doi: 10.1080/21505594.2017.1317412

Cejas, D., Canigia, L. F., Cruz, G. R., Elena, A. X., Maldonado, I., Gutkind, G. O., et al. (2014). First isolate of KPC-2-producing Klebsiella pneumonaie sequence type 23 from the Americas. *J. Clin. Microbiol.* 52, 3483–3485. doi: 10.1128/jcm.00726-14

Cejas, D., Elena, A., Guevara Nunez, D., Sevillano Platero, P., De Paulis, A., Magarinos, F., et al. (2019). Changing epidemiology of KPC-producing *Klebsiella pneumoniae* in Argentina: Emergence of hypermucoviscous ST25 and high-risk clone ST307. *J. Glob. Antimicrob. Resist*. 18, 238–242. doi: 10.1016/j.jgar.2019.06.005

Cubero, M., Grau, I., Tubau, F., Pallares, R., Dominguez, M. A., Linares, J., et al. (2016). Hypervirulent *Klebsiella pneumoniae* clones causing bacteraemia in adults in a teaching hospital in Barcelona. Spain (2007-2013). *Clin. Microbiol. Infect.* 22, 154–160. doi: 10.1016/j.cmi.2015.09.025

Dong, N., Lin, D., Zhang, R., Chan, E. W., and Chen, S. (2018a). Carriage of blaKPC-2 by a virulence plasmid in hypervirulent *Klebsiella pneumoniae*. *J. Antimicrob. Chemother.* 73, 3317–3321. doi: 10.1093/jac/dky358

Dong, N., Sun, Q., Huang, Y., Shu, L., Ye, L., Zhang, R., et al. (2019). Evolution of Carbapenem-Resistant Serotype K1 Hypervirulent *Klebsiella pneumoniae* by Acquisition of bla_{VIM-1}-Bearing Plasmid. *Antimicrob. Agents Chemother.* 63, e01056–e01019. doi: 10.1128/aac.01056-19

Dong, N., Yang, X., Zhang, R., Chan, E. W.C., and Chen, S. (2018b). Tracking microevolution events among ST11 carbapenemase-producing hypervirulent *Klebsiella pneumoniae* outbreak strains. *Emerg. Microb. Infect*. 7, 146–153. doi: 10.1038/s41426-018-0146-6

Durao, P., Balbontin, R., and Gordo, I. (2018). Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance. *Trends Microbiol.* 26, 677–691. doi: 10.1016/j.tim.2018.01.005

Feng, Y., Lu, Y., Yao, Z., and Zong, Z. (2018). Carbapenem-Resistant Hypervirulent *Klebsiella pneumoniae* of Sequence Type 36. *Antimicrob. Agents Chemother.* 62, e2644–e2617. doi: 10.1128/aac.02644-17

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

Fu, Y., Xu, M., Liu, Y., Li, A., and Zhou, J. (2019). Virulence and genomic features of a bla(CTX-M-3) and bla(CTX-M-14) coharboring hypermucoviscous *Klebsiella pneumoniae* of serotype K2 and ST65. *Infect. Drug Resist*. 12, 145–159. doi: 10.2147/idr.S187289

Gu, D. X., Huang, Y. L., Ma, J. H., Zhou, H. W., Fang, Y., Cai, J. C., et al. (2016). Detection of Colistin Resistance Gene mcr-1 in Hypervirulent *Klebsiella pneumoniae* and *Escherichia coli* Isolates from an Infant with Diarrhea in China. *Antimicrob. Agents Chemother.* 60, 5099–5100. doi: 10.1128/aac.00476-16

Gu, D., Dong, N., Zheng, Z., Lin, D., Huang, M., Wang, L., et al. (2017). A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. *Lancet Infect Dis*. 18, 37–46. doi: 10.1016/s1473-3099(17)30489-9

Gu, D., Lv, H., Sun, Q., Shu, L., and Zhang, R. (2018). Emergence of tet(A) and bla(KPC-2) co-carrying plasmid from a ST11 hypervirulent *Klebsiella pneumoniae* isolate in patient's gut. *Int. J. Antimicrob. Agents* 52, 307–308. doi: 10.1016/j.ijantimicag.2018.06.003

Harada, S., Aoki, K., Ishii, Y., Ohno, Y., Nakamura, A., Komatsu, M., et al. (2019). Emergence of IMP-producing hypervirulent *Klebsiella pneumoniae* carrying a pLVPK-like virulence plasmid. *Int. J. Antimicrob. Agents* 53, 873–875. doi: 10.1016/j.ijantimicag.2019.05.007

Huang, Y.-H., Chou, S.-H., Liang, S.-W., Ni, C.-E., Lin, Y.-T., Huang, Y.-W., et al. (2018). Emergence of an XDR and carbapenemase-producing hypervirulent *Klebsiella pneumoniae* strain in Taiwan. *J. Antimicrob. Chemother.* 73, 2039–2046. doi: 10.1093/jac/dky164

Lam, M. M. C., Wyres, K. L., Judd, L. M., Wick, R. R., Jenney, A., Brisse, S., et al. (2018). Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in *Klebsiella pneumoniae*. *Genome Med.* 10:77. doi: 10.1186/s13073-018-0587-5

Lam, M. M. C., Wyres, K. L., Wick, R. R., Judd, L. M., Fostervold, A., Holt, K. E., et al. (2019). Convergence of virulence and MDR in a single plasmid vector in MDR *Klebsiella pneumoniae* ST15. *J. Antimicrob. Chemother.* 74, 1218–1222. doi: 10.1093/jac/dkz028

Lee, C. R., Lee, J. H., Park, K. S., Jeon, J. H., Kim, Y. B., Cha, C. J., et al. (2017). Antimicrobial Resistance of Hypervirulent *Klebsiella pneumoniae*: Epidemiology. Hypervirulence-Associated Determinants, and Resistance Mechanisms. *Front Cell Infect Microbiol*. 7, 483–495. doi: 10.3389/fcimb.2017.00483

Lee, C. R., Lee, J. H., Park, K. S., Kim, Y. B., Jeong, B. C., and Lee, S. H. (2016). Global Dissemination of Carbapenemase-Producing *Klebsiella pneumoniae*: Epidemiology. Genetic Context, Treatment Options, and Detection Methods. *Front. Microbiol.* 7, 895–924. doi: 10.3389/fmicb.2016.00895

Li, J., Huang, Z.-Y., Yu, T., Tao, X.-Y., Hu, Y.-M., Wang, H.-C., et al. (2019). Isolation and characterization of a sequence type 25 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* from the mid-south region of China. *BMC Microbiol.* 19:219–219. doi: 10.1186/s12866-019-1593-5

Li, W., Sun, G. Z., Yu, Y. H., Li, N., Chen, M., Jin, R. H., et al. (2014). Increasing occurrence of antimicrobial-resistant hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* isolates in China. *Clin. Infect. Dis.* 58, 225–232. doi: 10.1093/cid/cit675

Lin, Y. T., Cheng, Y. H., Juan, C. H., Wu, P. F., Huang, Y. W., Chou, S. H., et al. (2018). High mortality among patients infected with hypervirulent antimicrobial-resistant capsular type K1 *Klebsiella pneumoniae* strains in Taiwan. *Int. J. Antimicrob. Agents* 52, 251–257. doi: 10.1016/j.ijantimicag.2018.06.008

Liu, B., and Su, W. (2019). Whole genome sequencing of NDM-1-producing serotype K1 ST23 hypervirulent *Klebsiella pneumoniae* in China. *J. Med. Microbiol.* 68, 866–873. doi: 10.1099/jmm.0.000996

Liu, P. P., Liu, Y., Wang, L. H., Wei, D. D., and Wan, L. G. (2016). Draft Genome Sequence of an NDM-5-Producing *Klebsiella pneumoniae* Sequence Type 14 Strain of Serotype K2. *Genom. Announc*. 4, e01610–e01615. doi: 10.1128/genomeA.01610-15

Liu, Y., Long, D., Xiang, T., Du, F., Wei, D. D., Wan, L. G., et al. (2019). Whole genome assembly and functional portrait of hypervirulent extensively drug-resistant NDM-1 and KPC-2 co-producing *Klebsiella pneumoniae* of capsular serotype K2 and ST86. *J. Antimicrob. Chemother.* 74, 1233–1240. doi: 10.1093/jac/dkz023

Liu, Z., Gu, Y., Li, X., Liu, Y., Ye, Y., Guan, S., et al. (2019). Identification and Characterization of NDM-1-producing Hypervirulent (Hypermucoviscous) *Klebsiella pneumoniae* in China. *Ann. Lab. Med.* 39, 167–175. doi: 10.3343/alm.2019.39.2.167

Lu, Y., Feng, Y., McNally, A., and Zong, Z. (2018). The Occurence of Colistin-Resistant Hypervirulent *Klebsiella pneumoniae* in China. *Front. Microbiol.* 9, 2568–2574. doi: 10.3389/fmicb.2018.02568

Million-Weaver, S., and Camps, M. (2014). Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? *Plasmid* 75, 27–36. doi: 10.1016/j.plasmid.2014.07.002

Navon-Venezia, S., Kondratyeva, K., and Carattoli, A. (2017). *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol. Rev.* 41, 252–275. doi: 10.1093/femsre/fux013

Pinilla-Redondo, R., Cyriaque, V., Jacquiod, S., Sorensen, S. J., and Riber, L. (2018). Monitoring plasmid-mediated horizontal gene transfer in microbiomes: recent advances and future perspectives. *Plasmid*. 99, 56–67. doi: 10.1016/j.plasmid.2018.08.002

Rafiq, Z., Sam, N., and Vaidyanathan, R. (2016). Whole genome sequence of *Klebsiella pneumoniae* U25, a hypermucoviscous, multidrug resistant, biofilm producing isolate from India. *Mem. Inst. Oswaldo Cruz*. 111, 144–146. doi: 10.1590/0074-02760150423

Ramsay, J. P., and Firth, N. (2017). Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* 38, 1–9. doi: 10.1016/j.mib.2017.03.003

Roulston, K. J., Bharucha, T., Turton, J. F., Hopkins, K. L., and Mack, D. J. F. (2018). A case of NDM-carbapenemase-producing hypervirulent *Klebsiella pneumoniae* sequence type 23 from the UK. *JMM Case Rep*. 5:e005130. doi: 10.1099/jmmcr.0.005130

Schweizer, C., Bischoff, P., Bender, J., Kola, A., Gastmeier, P., Hummel, M., et al. (2019). Plasmid-Mediated Transmission of KPC-2 Carbapenemase in *Enterobacteriaceae* in Critically Ill Patients. *Front. Microbiol.* 10:276. doi: 10.3389/fmicb.2019.00276

Shankar, C., Nabarro, L. E. B., Devanga Ragupathi, N. K., Muthuirulandi Sethuvel, D. P., Daniel, J. L. K., Doss, C. G. P., et al. (2016a). Draft Genome Sequences of Three Hypervirulent Carbapenem-Resistant *Klebsiella pneumoniae* Isolates from Bacteremia. *Genom. Announc*. 4, e1081–e1016. doi: 10.1128/genomeA.01081-16

Shankar, C., Santhanam, S., Kumar, M., Gupta, V., Devanga Ragupathi, N. K., and Veeraraghavan, B. (2016b). Draft Genome Sequence of an Extended-Spectrum-beta-Lactamase-Positive Hypervirulent *Klebsiella pneumoniae* Strain

with Novel Sequence Type 2318 Isolated from a Neonate. *Genom. Announc.* 4, e1273–e1216. doi: 10.1128/genomeA.01273-16

Shen, D., Ma, G., Li, C., Jia, X., Qin, C., Yang, T., et al. (2019). Emergence of a Multidrug-Resistant Hypervirulent *Klebsiella pneumoniae* Sequence Type 23 Strain with a Rare bla(CTX-M-24)-Harboring Virulence Plasmid. *Antimicrob. Agents Chemother.* 63, e2273–e2218. doi: 10.1128/aac.02273-18

Shon, A. S., Bajwa, R. P. S., and Russo, T. A. (2013). Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* A new and dangerous breed. *Virulence* 4, 107–118. doi: 10.4161/viru.22718

Shu, L., Dong, N., Lu, J., Zheng, Z., Hu, J., Zeng, W., et al. (2019). Emergence of OXA-232 Carbapenemase-Producing *Klebsiella pneumoniae* That Carries a pLVPK-Like Virulence Plasmid among Elderly Patients in China. *Antimicrob. Agents Chemother.* 63, e2246–e2218. doi: 10.1128/aac.02246-18

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74, 434–452. doi: 10.1128/MMBR.00020-10

Surgers, L., Boyd, A., Girard, P.-M., Arlet, G., and Decre, D. (2016). ESBL-Producing Strain of Hypervirulent Kiebsiella pneumoniae K2. *France. Emerg. Infect. Dis.* 22, 1687–1688. doi: 10.3201/eid2209.160681

Tabrizi, A. M. A., Badmasti, F., Shahcheraghi, F., and Azizi, O. (2018). Outbreak of hypervirulent *Klebsiella pneumoniae* harbouring bla(VIM-2) among mechanically-ventilated drug-poisoning patients with high mortality rate in Iran. *J. Glob. Antimicrob. Resist.* 15, 93–98. doi: 10.1016/j.jgar.2018.06.020

Turton, J. F., Payne, Z., Coward, A., Hopkins, K. L., Turton, J. A., Doumith, M., et al. (2017). Virulence genes in isolates of *Klebsiella pneumoniae* from the UK during 2016, including among carbapenemase gene-positive hypervirulent K1-ST23 and 'non-hypervirulent'types ST147, ST15 and ST383. *J. Med. Microbiol.* 67, 118–128. doi: 10.1099/jmm.0.000653

Turton, J., Davies, F., Turton, J., Perry, C., Payne, Z., and Pike, R. (2019). Hybrid Resistance and Virulence Plasmids in "High-Risk". Clones of *Klebsiella pneumoniae*, Including Those Carrying blaNDM-5. *Microorganisms* 7, 326–336. doi: 10.3390/microorganisms7090326

Tzouvelekis, L. S., Markogiannakis, A., Psichogiou, M., Tassios, P. T., and Daikos, G. L. (2012). Carbapenemases in *Klebsiella pneumoniae* and other *Enterobacteriaceae*: an evolving crisis of global dimensions. *Clin. Microbiol. Rev.* 25, 682–707. doi: 10.1128/CMR.05035-11

Wei, D. D., Wan, L. G., and Liu, Y. (2018). Draft Genome Sequence of an NDM-1- and KPC-2-Coproducing Hypervirulent Carbapenem-Resistant *Klebsiella pneumoniae* Strain Isolated from Burn Wound Infections. *Genom. Announc.* 6, e118–e192. doi: 10.1128/genomeA.00192-18

Wei, D. D., Wan, L. G., Deng, Q., and Liu, Y. (2016). Emergence of KPC-producing *Klebsiella pneumoniae* hypervirulent clone of capsular serotype K1 that belongs to sequence type 11 in Mainland China. *Diagn. Microbiol. Infect. Dis.* 85, 192–194. doi: 10.1016/j.diagmicrobio.2015.03.012

Wong, M. H. Y., Shum, H.-P., Chen, J. H. K., Man, M.-Y., Wu, A., Chan, E. W.-C., et al. (2018). Emergence of carbapenem-resistant hypervirulent *Klebsiella pneumoniae*. *Lancet Infect. Dis.* 18, 24–25. doi: 10.1016/s1473-3099(17)30629-1

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

Wyres, K. L., Wick, R. R., Judd, L. M., Froumine, R., Tokolyi, A., Gorrie, C. L., et al. (2019). Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet.* 15, e1008114–e1008138. doi: 10.1371/journal.pgen.1008114

Xie, M., Chen, K., Ye, L., Yang, X., Xu, Q., Yang, C., et al. (2020). Conjugation of Virulence Plasmid in Clinical *Klebsiella pneumoniae* Strains through Formation of a Fusion Plasmid. *Adv. Biosyst.* 4:e1900239. doi: 10.1002/adbi.201900239

Xie, Y., Tian, L., Li, G., Qu, H., Sun, J., Liang, W., et al. (2018). Emergence of the third-generation cephalosporin-resistant hypervirulent *Klebsiella pneumoniae* due to the acquisition of a self-transferable blaDHA-1-carrying plasmid by an ST23 strain. *Virulence* 9, 838–844. doi: 10.1080/21505594.2018.1456229

Xu, M., Fu, Y. Q., Fang, Y. H., Xu, H., Kong, H. S., Liu, Y. C., et al. (2019). High prevalence of KPC-2-producing hypervirulent *Klebsiella pneumoniae* causing meningitis in Eastern China. *Infect. Drug Resist.* 12, 641–653. doi: 10.2147/idr.S191892

Xu, M., Li, A., Kong, H. S., Zhang, W. L., Chen, H. C., Fu, Y. J., et al. (2018). Endogenous endophthalmitis caused by a multidrug-resistant hypervirulent *Klebsiella pneumoniae* strain belonging to a novel single locus variant of ST23: first case report in China. *BMC Infect. Dis.* 18, 669–674. doi: 10.1186/s12879-018-3543-5

Yao, B., Xiao, X. M., Wang, F., Zhou, L., Zhang, X. W., and Zhang, J. (2015). Clinical and molecular characteristics of multi-clone carbapenem-resistant hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* isolates in a tertiary hospital in Beijing. *China. Int. J. Infect. Dis.* 37, 107–112. doi: 10.1016/j.ijid.2015.06.023

Yao, H., Qin, S., Chen, S., Shen, J., and Du, X. D. (2018). Emergence of carbapenem-resistant hypervirulent *Klebsiella pneumoniae*. *Lancet Infect Dis.* 18:25. doi: 10.1016/s1473-3099(17)30628-x

Yuan, Y., Li, Y., Wang, G., Li, C., Chang, Y.-F., Chen, W., et al. (2019). bla(NDM-5) carried by a hypervirulent *Klebsiella pneumoniae* with sequence type 29. *Antimicrob. Resist. Infect. Control.* 8, 142–149. doi: 10.1186/s13756-019-0596-1

Zhan, L. L., Wang, S. S., Guo, Y. J., Jin, Y., Duan, J. J., Hao, Z. H., et al. (2017). Outbreak by Hypermucoviscous *Klebsiella pneumoniae* ST11 Isolates with Carbapenem Resistance in a Tertiary Hospital in China. *Front. Cell Infect. Microbiol.* 7, 182–190. doi: 10.3389/fcimb.2017.00182

Zhang, J., Zhou, K., Zheng, B. W., Zhao, L. N., Shen, P., Ji, J. R., et al. (2016). High Prevalence of ESBL-Producing *Klebsiella pneumoniae* Causing Community-Onset Infections in China. *Front. Microbiol.* 7, 1830–1839. doi: 10.3389/fmicb.2016.01830

Zhang, R., Lin, D. C., Chan, E. W. C., Gu, D. X., Chen, G. X., and Chen, S. (2015). Emergence of Carbapenem-Resistant Serotype K1 Hypervirulent *Klebsiella pneumoniae* Strains in China. *Antimicrob. Agents Chemother.* 60, 709–711. doi: 10.1128/aac.02173-15

Zhang, Y. W., Zeng, J., Liu, W. E., Zhao, F., Hu, Z. D., Zhao, C. J., et al. (2015). Emergence of a hypervirulent carbapenem-resistant *Klebsiella pneumoniae* isolate from clinical infections in China. *J. Infect.* 71, 553–560. doi: 10.1016/j.jinf.2015.07.010

Zhang, Y. W., Zhao, C. J., Wang, Q., Wang, X. J., Chen, H. B., Li, H. A., et al. (2016). High Prevalence of Hypervirulent *Klebsiella pneumoniae* Infection in China: Geographic Distribution, Clinical Characteristics, and Antimicrobial Resistance. *Antimicrob. Agents Chemother.* 60, 6115–6120. doi: 10.1128/aac.01127-16

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.

Copyright © 2020 Tang, Kong, Hao and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 25 November 2020
doi: 10.3389/fcimb.2020.614053

Identification of Uncommon *Cryptosporidium viatorum* (a Novel Subtype XVcA2G1c) and *Cryptosporidium andersoni* as Well as Common *Giardia duodenalis* Assemblages A and B in Humans in Myanmar

Yanchen Wu[1†], Baiyan Gong[1†], Xiaohua Liu[1], Yanyan Jiang[2], Jianping Cao[2], Lan Yao[1], He Li[1], Aiqin Liu[1] and Yujuan Shen[2*]*

[1] Department of Parasitology, Harbin Medical University, Harbin, China, [2] National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Chinese Center for Tropical Diseases Research, WHO Collaborating Centre for Tropical Diseases, National Center for International Research on Tropical Diseases, Ministry of Science and Technology, Key Laboratory of Parasite and Vector Biology, MOH, Shanghai, China

Cryptosporidium and *Giardia* are two important zoonotic intestinal protozoa responsible for diarrheal diseases in humans and animals worldwide. Feces from infected hosts, water and food contaminated by *Cryptosporidium* oocysts and *Giardia* cysts as well as predictors such as poverty have been involved in their transmission. Myanmar is one of the world's most impoverished countries. To date, there are few epidemiological studies of *Cryptosporidium* and *Giardia* in humans. To understand the prevalence and genetic characterization of *Cryptosporidium* spp. and *Giardia duodenalis* in humans in Myanmar, a molecular epidemiological investigation of the two protozoa was conducted in four villages of Shan State. 172 fecal specimens were collected from Wa people (one each) and identified for the presence of *Cryptosporidium* spp. and *G. duodenalis* by sequence analysis of their respective small subunit ribosomal RNA genes. 1.74% of investigated people were infected with *Cryptosporidium* spp.—*C. andersoni* (n = 2) and *C. viatorum* (n = 1) while 11.05% infected with *G. duodenalis*—assemblages A (n = 6) and B (n = 13). By sequence analysis of 60-kDa glycoprotein gene, the *C. viatorum* isolate belonged to a novel subtype XVcA2G1c. DNA preparations positive for *G. duodenalis* were further subtyped. Five of them were amplified and sequenced successfully: different assemblage B sequences (n = 2) at the triosephosphate isomerase (tpi) locus; sub-assemblage AII sequence (n = 1) and identical assemblage B sequences (n = 2) at the β-giardin (bg) locus. This is the first molecular epidemiological study of *Cryptosporidium* spp. and *G. duodenalis* in humans in Myanmar at both genotype and subtype levels. Due to unclear transmission patterns and dynamics of *Cryptosporidium* spp. and *G. duodenalis*, future research effort should focus on molecular epidemiological investigations of the two

OPEN ACCESS

Edited by:
Zhan Zhou,
Zhejiang University, China

Reviewed by:
Junqiang Li,
Henan University of Traditional Chinese Medicine, China
Meng Qi,
Tarim University, China

***Correspondence:**
Aiqin Liu
liuaiqin1128@126.com
Yujuan Shen
amyshyj12@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Parasite and Host, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *05 October 2020*
Accepted: *30 October 2020*
Published: *25 November 2020*

Citation:
Wu Y, Gong B, Liu X, Jiang Y, Cao J, Yao L, Li H, Liu A and Shen Y (2020) Identification of Uncommon Cryptosporidium viatorum (a Novel Subtype XVcA2G1c) and Cryptosporidium andersoni as Well as Common Giardia duodenalis Assemblages A and B in Humans in Myanmar. Front. Cell. Infect. Microbiol. 10:614053. doi: 10.3389/fcimb.2020.614053

parasites in humans and animals living in close contact in the investigated areas, even in whole Myanmar. These data will aid in making efficient control strategies to intervene with and prevent occurrence of both diseases.

Keywords: *Cryptosporidium*, *Giardia*, humans, genotyping, subtyping

INTRODUCTION

Cryptosporidium and *Giardia* are two ubiquitous intestinal protozoan parasites in humans and numerous animals. Both cryptosporidiosis and giardiasis are clinically characterized by diarrhea, and the severity of diarrhea is closely related to the age and health status of the infected hosts as well as the genetic background and infective dose of the parasites (Xiao and Fayer, 2008). Immunocompetent individuals typically experience self-limiting diarrhea and are often asymptomatic while chronic diarrhea in immunocompromised individuals (Plutzer et al., 2018). Severe life-threatening diarrhea has been reported in cryptosporidiosis patients infected with human immunodeficiency virus (HIV) (Ryan et al., 2016). The infective dose of the two parasites are low: <10 *Cryptosporidium* (*Cryptosporidium hominis* or *Cryptosporidium parvum*) oocysts and 10–100 *Giardia duodenalis* (syn. *Giardia intestinalis*, *Giardia lamblia*) cysts can cause infection in immunocompetent persons (Rendtorff, 1954; Okhuysen et al., 1999; Chappell et al., 2006). More seriously, a single *C. parvum* oocyst has been reported to initiate infection in immunosuppressed persons (Zhao et al., 2014). In general, the pathogenicity of *Cryptosporidium* is considered to be more severe in humans than that of *Giardia* (Prado et al., 2005). Humans can acquire *Cryptosporidium* and *Giardia* infections through the fecal-oral route, either directly (via human-to-human/animal contact) or indirectly (via ingestion of contaminated water or food) (Osman et al., 2016). The role of water and food in the epidemiology of the two parasitic diseases is now well recognized. To date, waterborne and foodborne outbreaks of cryptosporidiosis (>524 and >26) (Liu et al., 2020) and giardiasis (>344 and >38) had been reported worldwide (Karanis et al., 2007; Baldursson and Karanis, 2011; Efstratiou et al., 2017; Ryan et al., 2019). Based on clinical and public health importance, *Cryptosporidium* and *Giardia* are listed on the Environmental Protection Agency (EPA) microbial contaminant candidate list of concern for waterborne transmission (https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations). The two pathogens have also been ranked as the 5th and 11th most important foodborne parasites worldwide by a joint Food and Agriculture Organization (FAO)/World Health Organization (WHO) in 2014, respectively (Plutzer et al., 2018).

Both *Cryptosporidium* and *Giardia* are complicated genera. To date, 41 *Cryptosporidium* species and over 40 genotypes have been recognized (Feng et al., 2018; Holubová et al., 2019; Bolland et al., 2020; Holubová et al., 2020). Among them, 22 *Cryptosporidium* species/genotypes have been identified in humans (Xiao and Feng, 2017; Kváč et al., 2018), and *C. hominis* and *C. parvum* are the two most common species, reported in > 90% of human cryptosporidiosis cases (Squire and Ryan, 2017). However, in some countries, especially in developing countries, some unusual species have a high occurrence in human cryptosporididosis cases, such as *Cryptosporidium meleagridis* (10–20%) in Thailand and in Peru (Gatei et al., 2002; Cama et al., 2007; Cama et al., 2008); *Cryptosporidium andersoni* (79.59%) in India (Hussain et al., 2017); *Cryptosporidium viatorum* (7.14–11.11%) in Ethiopia (Adamu et al., 2014; de Lucio et al., 2016). Among the eight recognized species, and only *G. duodenalis* has been found to infect humans with eight assemblages (A to H) being identified (Ryan et al., 2019). Assemblages A and B are responsible for the vast majority (99%) of human giardiasis cases, and both of them have also been found in a variety of mammal species (Sprong et al., 2009). Assemblages C to H are specific to some animal species, but assemblages C to F are occasionally found in humans (Cacciò et al., 2018).

Cryptosporidiosis and giardiasis cause considerable human disease burdens worldwide. Like other infectious diseases, these two parasitic diseases also usually affect people living in poverty, and may further promote poverty. Developing countries usually have higher prevalence than developed countries: 5–10% versus 1% for cryptosporidiosis (Checkley et al., 2015); 0.9–40.7% versus 0.4–7.0% for giardiasis (Feng and Xiao, 2011; Ryan and Cacciò, 2013). In fact, early in 2004, *Cryptosporidium* spp. and *G. duodenalis* were included in the WHO's "Neglected Disease Initiative" due to their link with poverty (Savioli et al., 2006). Myanmar is one of the world's most impoverished countries. As of 2019, Myanmar ranks 145 out of 189 countries according to the Human Development Index (http://hdr.undp.org/en/content/2019-human-development-index-ranking). However, it is unclear on epidemiological status of human cryptosporidiosis and giardiasis and genetic characterization of *Cryptosporidium* spp. and *G. duodenalis*. To date, only two studies reported the prevalence by microscopy in Myanmar: *Cryptosporidium* spp. (3.4%, 7/203) in infants with diarrhea (Aye et al., 1994) and *G. duodenalis* (3.4%, 28/821) in schoolchildren and guardians (Kim et al., 2016). The present study was conducted to understand the prevalence and genetic characterization of *Cryptosporidium* spp. and *G. duodenalis* in humans in Myanmar at the genotype and subtype levels.

MATERIALS AND METHODS

Ethics Statement

Scientific approval and ethical clearance for this study was given by the Ethics Committee of the National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, China, and the Myanmar Eastern Shan State Special Region 2

Ethic Health Organization. All study participants were informed of the aims and the procedures of this study at enrollment. Before collection of fecal specimens, written informed consents for all study participants were obtained from each adult individual or legal guardian of enrolled children.

Study Site and Collection of Fecal Specimens

In October, 2018, a total of 172 fecal specimens (approximately 5–10 g each) were collected from Wa people (one specimen each) in four villages of Pangsang Township (22°10′N, 99°11′E) of Matman District of Shan State, which is located in the east of Myanmar, bordering with China's Yunnan Province in about 133-kilometer-long border line (**Figure 1**). The villages investigated have poor sanitation and garbage collection. Domestic pigs and chickens are the most common animals, which are kept by almost every household. The villagers are poorly educated and have weak hygiene awareness. The participants were composed of children (n = 97), teenagers (n = 41) and adults (n = 34), with their ages ranging from seven to 53 years. At the time of sampling, we only recorded the presence or absence of diarrhea. All fecal specimens were delivered to the laboratory in a cooler with ice packs within 24 h after collection and stored in a refrigerator at −20°C for future analysis.

DNA Extraction

Genomic DNA was extracted directly from approximately 180 to 200 mg of each fecal specimen using a QIAamp DNA Mini Stool Kit (Qiagen, Hilden, Germany) according to the manufacturer-recommended procedures. To obtain a high yield of DNA, the lysis temperature was increased to 95°C. DNA was eluted in 200 μl of AE elution buffer and stored at −20°C. Extracted DNA preparations were analyzed by nested polymerase chain reaction (PCR) amplification.

Genotyping and Subtyping of *Cryptosporidium* spp. and *G. duodenalis*

Cryptosporidium species were identified by nested PCR amplification of the partial small subunit ribosomal RNA (SSU rRNA) gene (approximately 830 bp) (Huang et al., 2016). Subtyping of DNA preparations positive for *Cryptosporidium* at the SSU rRNA locus was performed by nested PCR amplification of the partial 60 kDa glycoprotein (gp60) gene (approximately 950 bp) (Stensvold et al., 2015). Meanwhile, all DNA preparations were screened for the presence of *G. duodenalis* by nested PCR amplification of the partial SSU rRNA gene (approximately 290 bp) and were identified to the assemblage level as previously described by Appelbee et al. (2003). DNA preparations positive for *G. duodenalis* at the SSU rRNA locus were further analyzed to determine sub-assemblages by nested PCR amplification of the triose phosphate isomerase (tpi) and β-giardin (bg) genes. Assemblage A/B-specific nested PCRs were performed to amplify approximately 330 and 460 bp nucleotide fragments of the tpi gene, respectively (Geurden et al., 2008; Levecke et al., 2009). Approximately 510 bp fragment of bg gene was amplified (Lalle et al., 2005).

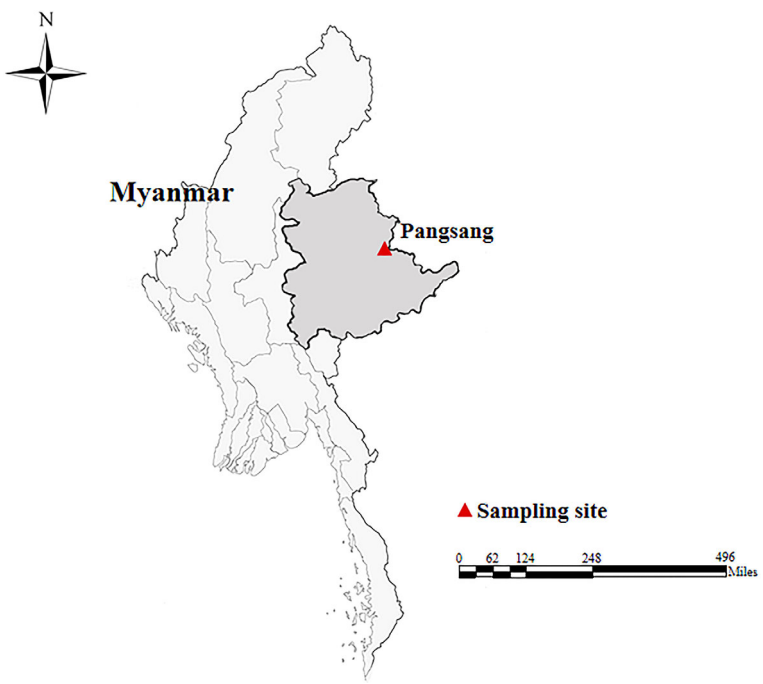


FIGURE 1 | A map showing the sampling site of human fecal specimens in this study.

Each DNA preparation was performed two times and TaKaRa Taq DNA polymerase (TaKaRa Bio Inc., Tokyo, Japan) was used for all PCR reactions. A negative control (DNase-free water) and a positive control (*C. baileyi* or *G. duodenalis* assemblage E) were included in all PCR tests. All secondary PCR products were subjected to electrophoresis in a 1.5% agarose gel and visualized by staining the gel with GelStrain (TransGen Biotech., Beijing, China) before sequencing.

Sequence Analysis

Positive secondary PCR products of expected size were sent to Comate Bioscience Company Limited (Jilin, China) for sequencing using their respective secondary PCR primers on an ABI PRISM 3730 XL DNA Analyzer using the BigDye Terminator v3.1Cycle Sequencing Kit (Applied Biosystems, Carlsbad, CA, USA). The accuracy of the sequencing data was confirmed by sequencing in both directions. Species/genotypes and subtypes of *Cryptosporidium*, and assemblages and sub-assemblages of *G. duodenalis* were identified by comparing the nucleotide sequences obtained in the present study with reference sequences downloaded from GenBank using the Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/blast/) and Clustal X 1.83 (http://www.clustal.org/).

Phylogenetic Analysis

To assess phylogenetic relationships among *C. viatorum* subtypes obtained in the present study and those published in GenBank databases, all gp60 gene sequences of *C. viatorum* subtypes were implemented in the software Mega 5 (http://www.megasoftware.net/). A neighbor-joining tree was constructed based on the evolutionary distances calculated by the Kimura 2-parameter model. The reliability of the trees was assessed using the bootstrap analysis with 1,000 replicates.

Statistical Analysis

All statistical analyses were performed with Statistical Package for the Social Sciences (SPSS) 19.0. Pearson chi-square (χ^2) and Fisher's exact tests were used to determine statistical significance in the present study. All results were interpreted using odds ratios, 95% confidence intervals and significance level (*P*-values < 0.05).

Nucleotide Sequence Accession Numbers

The representative nucleotide sequences obtained in this study were deposited in the GenBank database under the following accession numbers: MW014313 to MW014315 (SSU rRNA) and MW014316 (gp60) for *Cryptosporidium*; MW011715 and MW011716 (tpi), MW011717 and MW011718 (bg) for *G. duodenalis*.

RESULTS

Prevalence of *Cryptosporidium* spp. and *G. duodenalis*

All 172 fecal specimens were screened for the presence of *Cryptosporidium* spp. and *G. duodenalis* by PCR amplification and sequence analysis of their respective partial SSU rRNA gene. *G. duodenalis* (11.05%, 19/172) was observed to be more prevalent than *Cryptosporidium* spp. (1.74%, 3/172) in the investigated people. *Cryptosporidium* spp. was found only in children (1.03%, 1/97) and adults (5.88%, 2/34) while *G. duodenalis* only in children (12.37%, 12/97) and teenagers (17.07%, 7/41) (**Table 1**).

By χ^2 tests, only a statistically higher prevalence of *G. duodenalis* was observed in children than in adults ($P = 0.04$). Both prevalences of *Cryptosporidium* spp. and *G. duodenalis* were higher in people without diarrhea than those with diarrhea (**Table 2**). Furthermore, there were no relationships between *Cryptosporidium* spp. or *G. duodenalis* infection and diarrhea in each age group ($P > 0.05$) (**Table 3**).

Genotyping and Subtyping of *Cryptosporidium*

Sequence analysis of the SSU rRNA gene identified two *Cryptosporidium* species: *C. andersoni* (n = 2) and *C. viatorum* (n = 1). Two *C. andersoni* isolates (MW014313 and MW014314) had 99.73% (two-base difference) and 99.62% (three-base difference) similarity with *C. andersoni* isolates from a dairy calf (JX515549), respectively. *C. viatorum* isolate was identical to a wild rat-derived *C. viatorum* isolate (MK522269).

TABLE 1 | Prevalence and genetic characterization of *Cryptosporidium* spp. and *G. duodenalis* in humans.

Age group (years)	Examined no.	*Cryptosporidium* spp.			*G. duodenalis*			
		No. of positive (%)	Species	Subtype	No. of positive (%)	Assemblage	Sub-assemblage	
			SSU rRNA (n)	Gp60 (n)		SSU rRNA (n)	tpi (n)	bg (n)
Children (<13)	97	1 (1.03)	*C. viatorum* (1)	XVcA2G1c (1)	12 (12.37)	B (9); A (3)	B-I (1)	B (2); AII (1)
Teenagers (13–17)	41	0	–	–	7 (17.07)	B (4); A (3)	B-II (1)	–
Adults (≥18)	34	2 (5.88)	*C. andersoni* (2)	–	0	–	–	–
Total	172	3 (1.74)	*C. viatorum* (1); *C. andersoni* (2)	XVcA2G1c (1)	19 (11.05)	B (13); A(6)	B-I (1); B-II (1)	B (2); AII (1)

The bars denote negative results at the locus.

TABLE 2 | Prevalence of *Cryptosporidium* spp. and *G. duodenalis* by age and symptom.

Group		Examined no.	*Cryptosporidium* spp.			*G. duodenalis*		
			Positive no. (%)	OR[a] (95% CI[b])	χ^2/*P*-value	Positive no. (%)	OR[a] (95% CI[b])	χ^2/*P*-value
Age	Children	97	1 (1.03)	Ref		12 (12.37)	Ref	
	Teenagers	41	0	0.99 (0.97, 1.01)	–/1.00[c]	7 (17.07)	0.69 (0.25, 1.89)	0.54/0.46
	Adults	34	2 (5.88)	0.17 (0.02, 1.90)	0.92/0.34	0	0.88 (0.81, 0.94)	**–/0.04**[c,d]
Symptom	Diarrhea	42	0	1.02 (1.00, 1.05)	–/1.00[c]	4 (9.52)	0.81 (0.25, 2.58)	0.01/0.94
	Non-diarrhea	130	3 (2.31)			15 (11.54)		

[a]*OR Odds ratio.* [b]*CI Confidence interval.* [c]*Fisher's exact test.* [d]*Bold type for values indicates statistical significance.*

TABLE 3 | Relationships between *Cryptosporidium* spp. or *G. duodenalis* infection and diarrhea in each age group.

Group	Symptom	Examined no.	*Cryptosporidium* spp.			*G. duodenalis*		
			Positive no. (%)	OR[a] (95% CI[b])	χ^2/*P*-value	Positive no. (%)	OR[a] (95% CI[b])	χ^2/*P*-value
Children	Diarrhea	26	0	1.01 (0.99, 1.04)	–/1.00[c]	2 (7.69)	0.51 (0.10, 2.49)	0.25/0.62
	Non-diarrhea	71	1 (1.41)			10 (14.08)		
Teenagers	Diarrhea	14	0	–	–	2 (14.29)	0.73 (0.12, 4.37)	–/1.00
	Non-diarrhea	27	0			5 (18.52)		
Adults	Diarrhea	2	0	1.07 (0.96, 1.17)	–/1.00[c]	0	–	–
	Non-diarrhea	32	2 (6.25)			0		

[a]*OR Odds ratio,* [b]*CI Confidence interval,* [c]*Fisher's exact test.*

In phylogenetic analysis of the gp60 gene sequences, the *C. viatorum* isolate obtained in the present study was grouped with *C. viatorum* subtypes XVcA2G1b, XVcA2G1a, and XVcA2G1 (**Figure 2**). Meanwhile, this result was also supported by evolutionary analysis at the nucleotide and amino acid levels: low genetic variations (0.12–0.99% and 0.39–1.96%) compared to the other three subtypes in XVc subtype family; high genetic variations (2.13–21.47% and 3.59–35.60%) compared to all 11 subtypes in subtype families (XVa, XVb, XVd) (**Table 4**). According to the terminology of *C. viatorum* subtypes established by Stensvold et al. (2015), a

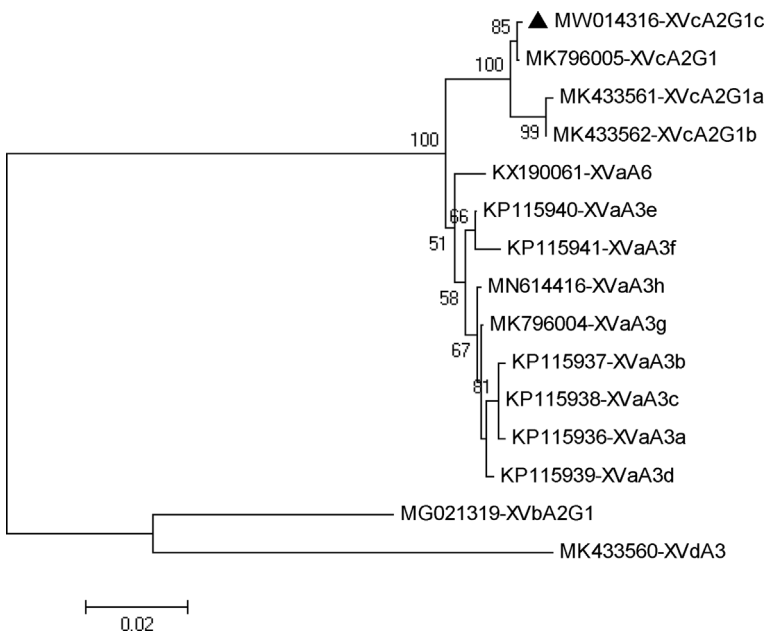


FIGURE 2 | Phylogenetic relationship of gp60 subtypes of *Cryptosporidium viatorum*. The relationships among *C. viatorum* subtypes identified in the present study and those deposited in the GenBank were inferred by a neighbor-joining analysis of gp60 gene sequences based on genetic distance by the Kimura 2-parameter model. The numbers on the branches are percent bootstrapping values from 1000 replicates. Each sequence is identified by its accession number and subtype designation. The triangle filled in black indicates the subtype identified in this study.

TABLE 4 | Pairwise differences (percentage) among gp60 subtypes of *C. viatorum* for nucleotide (below the diagonal) and amino acid sequences (above the diagonal).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1	–	0	0	0.79	0.79	2.38	0.40	0.79	3.16	28.40	4.40	5.20	5.20	4.40	33.20	1	XVaA3a
2	0.25	–	0	0.79	0.79	2.38	0.40	0.79	3.16	28.40	4.40	5.20	5.20	4.40	33.20	2	XVaA3b
3	0.12	0.12	–	0.79	0.79	2.38	0.40	0.79	3.16	28.40	4.40	5.20	5.20	4.40	33.20	3	XVaA3c
4	0.50	0.50	0.37	–	0.79	2.39	0.40	0.79	3.17	28.92	4.42	5.22	5.22	4.42	34.14	4	XVaA3d
5	1.00	1.00	0.87	0.75	–	1.58	0.39	0	2.75	28.97	3.97	4.76	4.76	3.97	33.73	5	XVaA3e
6	1.50	1.50	1.38	1.25	0.49	–	1.98	1.58	2.77	29.60	4.00	4.80	4.80	4.00	33.60	6	XVaA3f
7	0.49	0.49	0.37	0.25	0.62	1.12	–	0.39	2.76	29.08	3.98	4.78	4.78	3.98	33.86	7	XVaA3g
8	0.62	0.62	0.49	0.37	0.49	0.99	0.12	–	2.36	29.08	3.59	4.38	4.38	3.59	33.86	8	XVaA3h
9	1.62	1.62	1.50	1.63	1.24	1.24	1.49	1.37	–	28.85	3.95	5.53	5.53	3.95	32.81	9	XVaA6
10	16.81	16.97	16.97	17.04	17.54	17.78	17.53	17.69	17.96	–	30.80	31.20	31.20	30.80	23.83	10	XVbA2G1
11	2.53	2.53	2.40	2.28	2.13	2.14	2.39	2.26	2.25	17.62	–	1.57	1.57	0.39	35.60	11	XVcA2G1
12	3.18	3.18	3.05	2.93	2.52	2.53	2.78	2.65	2.90	17.93	0.86	–	0	1.96	36.00	12	XVcA2G1a
13	3.31	3.31	3.18	3.06	2.65	2.66	2.91	2.78	3.03	18.10	0.99	0.12	–	1.96	36.00	13	XVcA2G1b
14	2.66	2.66	2.53	2.41	2.13	2.14	2.39	2.26	2.25	17.79	0.12	0.86	0.99	–	35.60	14	**XVcA2G1c**
15	19.55	19.57	19.39	19.83	20.69	20.60	20.32	20.49	19.71	12.61	21.29	21.97	22.16	21.47	–	15	XVdA3
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		

The bolded subtype was obtained in the present study.

novel subtype XVcA2G1c was identified, which had the largest similarity (99.88%) with subtype XVcA2G1 (MK796005) from a Bower's white-toothed rat in China.

Genotyping and Subtyping of *G. duodenalis*

Sequence analysis of the SSU rRNA gene identified two *G. duodenalis* assemblages: A (n = 6) and B (n = 13). All assemblage A isolates were identical to each other, and so were assemblage B isolates, which had 100% similarity with two horse-derived isolates (MN174121 and MN174122), respectively.

At the tpi locus, only two isolates of assemblage B were successfully amplified and sequenced, with two different tpi gene sequences being obtained. Due to no clear subgrouping within assemblage B, both of them were named as sub-assemblage B-I (MW011715) and B-II (MW011716) for convenient description, respectively (**Table 1**). Sub-assemblage B-I had one-base difference compared to the JX994251 and KM977638 sequences from a human and a chinchilla, respectively. Sub-assemblage B-II had 100% similarity with those isolates from a human (JX994245) and animals—non-human primates (MK982533), a bos indicus (MF459680), a dog (LC437486), a cat (LC341576), a goat (MF095053), a pig (MH644772), a rabbit (MH475909), a fox (KY304077), an orangutan (KR011753), a chinchilla (KF843914), and an anteater (GU797247).

At the bg locus, only three *G. duodenalis* isolates were successfully amplified and sequenced. The assemblage A isolate was identified as sub-assemblage AII, which was identical to those *G. duodenalis* isolates from a human (MN844143), a sheep (MK452883), and a cattle (MK452836). Two assemblage B isolates were identical to each other, and shared 100% similarity with a human-derived *G. duodenalis* isolate (MK982542).

DISCUSSION

To the best of our knowledge, the present study is the first report of the prevalence and genetic characterization of *Cryptosporidium* spp. and *G. duodenalis* by molecular techniques in Wa people in Myanmar. *Cryptosporidium* spp. was detected in children (1.03%, 1/97) and adults (5.88%, 2/34), while *G. duodenalis* in children (12.37%, 12/97) and teenagers (17.07%, 7/41). However, in two studies conducted in Myanmar, a prevalence of 3.4% was found in detection of either *Cryptosporidium* spp. in infants or *G. duodenalis* in schoolchildren and guardians by microscopy (Aye et al., 1994; Kim et al., 2016). The difference in prevalence may be related to detection methods employed. PCR-based molecular techniques are demonstrated to be more sensitive than conventional microscopy, such as *Cryptosporidium* spp. prevalence in sheep in Australia (26.25% versus 2.6%) (Ryan et al., 2005) and in the US (50.8% versus 20.6%) (Santín et al., 2007); *G. duodenalis* prevalence in dogs in India (20.0% versus 3.0%) (Traub et al., 2004) and in Italy (20.5% versus 11.0%) (Scaramozzino et al., 2009). Besides that, the prevalence may be related to population specimens collected and their clinical features. Children are reported to have a statistically significantly higher prevalence than adults: such as 2.56% versus 1.89% for *Cryptosporidium* spp. in China (Liu et al., 2020); 53.2% versus 22.2% for *G. duodenalis* in Uganda (Johnston et al., 2010). Meanwhile, some studies reported significant difference in prevalence of both parasites between diarrheal and non-diarrheal children, such as 16.3% versus 3.1% for *Cryptosporidium* spp. in Tanzania (Tellevik et al., 2015); 20.5% versus 8.0% for *G. dudenalis* in Ethiopia (Feleke et al., 2018). However, in the present study, *G. duodenalis* was more prevalent in non-diarrheal cases than in diarrheal cases (14.08% versus 7.69% for children; 18.52% versus 14.29% for teenagers), and all three *Cryptosporidium*-positive cases were from non-diarrheal individuals. Similar results are also reported in some previous studies, such as *Cryptosporidium* spp. in Peru and *G. duodenalis* in Ethiopia (Cama et al., 2008; Tellevik et al., 2015). Even so, the two parasites can also lead to growth and development retardation of asymptomatic children (Prado et al., 2005; Checkley et al., 2015), thus constituting a serious public health problem in this population. The prevalences are complicated and difficult to compare due to differences in the detection methods employed, the size of specimens analyzed, the populations investigated, and the health status of individuals.

C. parvum and *C. hominis* are recognized as the main causative agents (> 90%) in reported human cases of *Cryptosporidium* infection (Squire and Ryan, 2017; Liu et al., 2020). However, in the present study, two unusual *Cryptosporidium* species were identified: *C. andersoni* (n = 2) and *C. viatorum* (n = 1). Since the first report of *C. andersoni* in humans in 2001, there have been 144 human cases of cryptosporidiosis attributed to *C. andersoni*, composed of 141 diarrheal cases from India (n = 78) (Hussain et al., 2017), China (n = 59) (Liu et al., 2014; Jiang et al., 2014; Su et al., 2017), the UK (n = 3) (Leoni et al., 2006) and Malawi (n = 1) (Morse et al., 2007), and three cases having no information on clinical symptoms from Australia (Waldron et al., 2011), Iran (Agholi et al., 2013), and France (Guyot et al., 2001). *C. viatorum* was first identified in 10 travellers with gastrointestinal symptoms returning to the UK from the Indian subcontinent in 2012 (Elwin et al., 2012). The name was chosen to underscore its link to foreign travel. To date, 37 human cases of cryptosporidiosis including the present case caused by *C. viatorum* have been identified in nine countries: Australia (n = 1) (Braima et al., 2019), China (n = 1) (Xu et al., 2020), Colombia (n = 1) (Sánchez et al., 2017), Ethiopia (n = 12) (Adamu et al., 2014; Stensvold et al., 2015; de Lucio et al., 2016), India (n = 2) (Khalil et al., 2017; Khalil et al., 2018), Myanmar (n = 1), Nigeria (n = 2) (Ayinmode et al., 2014; Ukwah et al., 2017), Sweden (n = 3) (Insulander et al., 2013; Stensvold et al., 2015), and the UK (n = 14) (Elwin et al., 2012; Stensvold et al., 2015). 45.95% (17/37) individuals had a history of travel abroad (Kenya, Guatemala, India, Barbados, Pakistan, Nepal and Bangladesh) (**Table 5**). Travel abroad was considered to be significantly associated with an increased risk of *Cryptosporidium* infections in the US and UK studies (Hunter et al., 2004; Roy et al., 2004). Meanwhile, it was observed that 75.68% (28/37) of human cases experienced diarrhea while 13.51% (5/37) had no diarrhea (**Table 5**). Occurrence of diarrhea and the severity of cryptosporidiosis in humans are complicated, often involving the immune status of the infected hosts, the virulence of *Cryptosporidium* species/genotypes, the infective dose of oocysts and other intestinal pathogens.

Currently, it is unclear on the source of infection/contamination of *C. andersoni* and *C. viatorum* in the investigated areas. *C. andersoni* is actually the major species causing cattle cryptosporidiosis, especially in yearlings and adults (Wang K. et al., 2019). With the accumulation of molecular epidemiological data of *Cryptosporidium*, *C. andersoni* has also been found occasionally in other animal species, such as sheep, horses, camels, golden takins, monkeys, hamsters, and ostriches (Liu et al., 2020). In a previous molecular epidemiological study of *Cryptosporidium* in diarrheal outpatients conducted in China, 21 C. *andersoni* isolates were identical to cattle/goat-derived isolates at the SSU rRNA locus (Jiang et al., 2014). In the present study, we observed high homology (99.73% and 99.62%) of the SSU rRNA gene of two *C. andersoni* isolates with a cattle-derived isolate. *C. viatorum* was initially thought to occur exclusively in humans. However, it has also been detected in some rat species in Australia (Koehler et al., 2018) and China (Chen et al., 2019; Zhao et al., 2019). The same subtypes (XVaA3h and XVaA3g) have been identified in humans (Braima et al., 2019; Xu et al., 2020) and rats (Chen et al., 2019) (**Tables 5**, **6**). In the present study, a novel subtype (XVcA2G1c) was identified in humans for the first time, which had the largest similarity of 99.88% (one-base difference) with that (XVcA2G1) from a Bower's white-toothed rat in China (Chen et al., 2019). These results above indicated the large potential of zoonotic transmission of *C. andersoni* and *C. viatorum*. Therefore, the true burden of

TABLE 5 | Geographical distribution of *C. viatorum* subtypes identified in humans and their travel history.

Host	Country/Travel history	Case number		Subtype (n)	Reference
		Diarrhea	Non-diarrhea		
Human	Australia/NA		1[a]	XVaA3g (1)	Braima et al., 2019
	China/NA		1	XVaA3h (1)	Xu et al., 2020
	Colombia/NA		1[a]	–	Sánchez et al., 2017
	Ethiopia/NA	8	2; 2[a]	XVaA3d (9)	Adamu et al., 2014; Stensvold et al., 2015; de Lucio et al., 2016
	India/NA	2		–	Khalil et al., 2017; Khalil et al., 2018
	Myanmar/NA		1	**XVcA2G1c (1)**	This study
	Nigeria/NA	2		–	Ayinmode et al., 2014; Ukwah et al., 2017
	Sweden/Kenya; Guatemala	3		XVaA3d (1); XVaA3c (1); XVaA3b (1)	Insulander et al., 2013; Stensvold et al., 2015
	UK/India; Barbados; Pakistan; Nepal; Bangladesh	13	1	XVaA3a (9), XVaA3f (2); XVaA3d (2); XVaA3e (1)	Elwin et al., 2012; Stensvold et al., 2015
Total		28	5; 4[a]	XVaA3a (9),XVaA3b (1); XVaA3c (1);XVaA3d (12); XVaA3f (2); XVaA3e (1); XVaA3h (1); XVaA3g (1); XVcA2G1c (1)	

NA, not available.
The bars denote negative results at the locus.
The bolded subtype was obtained in the present study.
[a]*Cases having no information on clinical symptoms.*

TABLE 6 | Geographical distribution of *C. viatorum* subtypes in animals.

Host	Country	Case number	Subtype (n)				Reference
			XVa	XVb	XVc	XVd	
Australian swamp rats	Australia	3		XVbA2G1b (3)			Koehler et al., 2018
Edward's long-tailed rats	China	11			XVcA2G1a (4); XVcA2G1b (1)	XVdA3 (1)	Zhao et al., 2019
	China	4	XVaA6 (2)				Chen et al., 2019
Bower's white-toothed rats	China	21	XVaA3g (7); XVaA3h (7)		XVcA2G1 (1)		Chen et al., 2019
Total		39	XVaA6 (2); XVaA3g (7); XVaA3h (7)	XVbA2G1b (3)	XVcA2G1a (4); XVcA2G1b (1); XVcA2G1 (1)	XVdA3 (1)	

human cryptosporidiosis caused by *C. andersoni* and *C. viatorum* attributed to humans and animals as well as the transmission dynamic of this disease needs to be assessed in the investigated areas by systematic molecular epidemiological surveys of humans and animals in the future.

The genotyping results showed all *G. duodenalis*-positive individuals in the investigated areas were infected with assemblages A (6/19, 31.58%) and B (13/19, 68.42%). Molecular epidemiological data indicated that assemblage B commonly had higher prevalence than assemblage A worldwide (Ryan and Cacciò, 2013), such as 68.0% (66/97) versus 29.9% (29/97) in Spain (Wang Y. et al., 2019); 66.7% (8/12) versus 33.3% (4/12) in China (Yu et al., 2019). However, an oppositive pattern of distribution of assemblages A and B has also been noticed in humans. Assemblage A showed a predominance compared to assemblage B in some studies, such as 35.9% (33/92) versus 21.7% (20/92) in Ethiopia (Damitie et al., 2018); 52.5% (31/59) versus 22.0% (13/59) in Ethiopia (Gelanew et al., 2007). Although epidemiological investigations have been conducted worldwide, the number of molecular epidemiological studies of giardiasis in humans is relatively small. To date, there is a lack of geographical structuring of the *G. duodenalis* assemblages across the globe. Difference in geographical distribution of assemblages might be associated to the socioeco-epidemiological factors of the population investigated (Sánchez et al., 2017) and methodological aspects, such as targeted genes, number of loci, primers, downstream procedures etc (Oliveira-Arbex et al., 2016).

One of six assemblage A isolates was successfully amplified and identified as sub-assemblage AII based on sequence analysis of the bg gene. Currently, there are three sub-assemblages (AI, AII and AIII) identified within assemblage A. Sub-assemblages AI and AII are commonly found in animals and humans, respectively; however, so far sub-assemblage AIII has been found only in animals, mostly in wildlife (Ryan et al., 2019). Although some studies indicated that contact with farm animals was associated with an increased risk of *G. duodenalis* infection for adults (Hoque et al., 2002; Hoque et al., 2003), genotyping and subtyping data point only to the potential role for zoonotic transmission with little epidemiological support (Xiao and Fayer, 2008). Sub-assemblage AII has ever been found in both pets (dogs) and their owners in Belgium (Claerebout et al., 2009). In India, genetically similar sub-assemblage AII isolates have been found in dogs and humans living within the same household (Traub et al., 2004). The same bg gene sequences of sub-assemblage AII were observed in human (here) and in sheep and cattle (previously). Meanwhile, at the tpi locus, one assemblage B isolate had 100% similarity with those from various animals (seen in Results). The finding of the same gene sequences of *G. duodenalis* isolates derived from humans and animals indicated the possibility of zoonotic transmission in the investigated areas. Due to the lack of data of *G. duodenalis* in local animals, the epidemiologic role of animals in the spread of giardiasis will be assessed.

CONCLUSION

This is the first molecular epidemiological investigation of *Cryptosporidium* spp. and *G. duodenalis* in humans in Myanmar. In general, *G. duodenalis* was more prevalent than *Cryptosporidium* spp. in the investigated areas. High percentage of non-diarrheal individuals infected with *Cryptosporidium* spp. and *G. duodenalis* should be made aware of the importance and epidemiological significance. Two unusual *Cryptosporidium* species (*C. andersoni* and *C. viatorum*) were identified, with a novel *C. viatorum* subtype XVcA2G1c being found for the first time. DNA sequences of *Cryptosporidium* spp. and *G. duodenalis* isolates from investigated people had high similarity or even identity of animal-derived isolates, implying the potential of zoonotic transmission. Due to unclear transmission patterns and dynamics of *Cryptosporidium* spp. and *G. duodenalis*, future research effort should focus on molecular epidemiological investigations of the two parasites in humans and animals living in close contact in the investigated areas, even in whole Myanmar. These data will aid in making efficient control strategies to intervene with and prevent occurrence of both diseases.

DATA AVAILABILITY STATEMENT

The representative nucleotide sequences obtained in this study were deposited in the GenBank database under the following accession numbers: MW014313 to MW014315 (SSU rRNA) and MW014316 (gp60) for *Cryptosporidium*; MW011715 and MW011716 (tpi), MW011717 and MW011718 (bg) for *G. duodenalis*.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics Committee of the National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, China, and the Myanmar Eastern Shan State Special Region 2 Ethic Health Organization. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AL and YS designed this study. YW, BG, and HL performed the experiments. YW, BG, XL, YJ, and LY analyzed the data. JC contributed reagents/materials. YW and BG wrote the first draft of the manuscript and prepared the tables and the figures. AL and YS made the final revision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported partially by the National Science and Technology Major Program of China (No. 2018ZX10713001-004 to YS), the National Key Research and Development Program of China (No. 2017YFD0501300 to YJ), the Fifth Round of Three-Year Public Health Action Plan of Shanghai (No. GWV-10.1-XK13 to JC).

ACKNOWLEDGMENTS

We thank the staff at the department of helminth, Yunnan Institute of Parasitic Diseases and the Centre for Disease Control and Prevention, and Zhiliang Zheng at Health Without Borders, for their assistance with sample collection and case investigation.

REFERENCES

Adamu, H., Petros, B., Zhang, G., Kassa, H., Amer, S., Ye, J., et al. (2014). Distribution and clinical manifestations of *Cryptosporidium* species and subtypes in HIV/AIDS patients in Ethiopia. *PLoS Negl. Trop. Dis.* 8, e2831. doi: 10.1371/journal.pntd.0002831

Agholi, M., Hatam, G. R., and Motazedian, M. H. (2013). HIV/AIDS-associated opportunistic protozoal diarrhea. *AIDS Res. Hum. Retroviruses* 29, 35–41. doi: 10.1089/AID.2012.0119

Appelbee, A. J., Frederick, L. M., Heitman, T. L., and Olson, M. E. (2003). Prevalence and genotyping of *Giardia duodenalis* from beef calves in Alberta, Canada. *Vet. Parasitol.* 112, 289–294. doi: 10.1016/s0304-4017(02)00422-3

Aye, T., Moe, K., Nyein, M. M., and Swe, T. (1994). Cryptosporidiosis in Myanmar infants with acute diarrhea. *Southeast Asian J. Trop. Med. Public Health* 25, 654–656.

Ayinmode, A. B., Zhang, H., Dada-Adegbola, H. O., and Xiao, L. (2014). *Cryptosporidium hominis* subtypes and *Enterocytozoon bieneusi* genotypes in HIV-infected persons in Ibadan, Nigeria. *Zoonoses Public Health* 61, 297–303. doi: 10.1111/zph.12072

Baldursson, S., and Karanis, P. (2011). Waterborne transmission of protozoan parasites: review of worldwide outbreaks-an update 2004–2010. *Water Res.* 45, 6603–6614. doi: 10.1016/j.watres.2011.10.013

Bolland, S. J., Zahedi, A., Oskam, C., Murphy, B., and Ryan, U. (2020). *Cryptosporidium bollandi* n. sp. (Apicomplexa: Cryptosporidiiae) from angelfish (*Pterophyllum scalare*) and Oscar fish (*Astronotus ocellatus*). *Exp. Parasitol.* 217, 107956. doi: 10.1016/j.exppara.2020.107956

Braima, K., Zahedi, A., Oskam, C., Reid, S., Pingault, N., Xiao, L., et al. (2019). Retrospective analysis of *Cryptosporidium* species in Western Australian human populations, (2015–2018), and emergence of the *C. hominis* IfA12G1R5 subtype. *Infect. Genet. Evol.* 73, 306–313. doi: 10.1016/j.meegid.2019.05.018

Cacciò, S. M., Lalle, M., and Svärd, S. G. (2018). Host specificity in the *Giardia duodenalis* species complex. *Infect. Genet. Evol.* 66, 335–345. doi: 10.1016/j.meegid.2017.12.001

Cama, V. A., Ross, J. M., Crawford, S., Kawai, V., Chavez-Valdez, R., Vargas, D., et al. (2007). Differences in clinical manifestations among *Cryptosporidium* species and subtypes in HIV-infected persons. *J. Infect. Dis.* 196, 684–691. doi: 10.1086/519842

Cama, V. A., Bern, C., Roberts, J., Cabrera, L., Sterling, C. R., Ortega, Y., et al. (2008). *Cryptosporidium* species and subtypes and clinical manifestations in children, Peru. *Emerg. Infect. Dis.* 14, 1567–1574. doi: 10.3201/eid1410.071273

Chappell, C. L., Okhuysen, P. C., Langer-Curry, R., Widmer, G., Akiyoshi, D. E., Tanriverdi, S., et al. (2006). *Cryptosporidium hominis*: experimental challenge of healthy adults. *Am. J. Trop. Med. Hyg.* 75, 851–857.

Checkley, W., White, A. C., Jaganath, D., Arrowood, M. J., Chalmers, R. M., Chen, X.-M., et al. (2015). A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*. *The Lancet Infect. Dis.* 15, 85–94. doi: 10.1016/s1473-3099(14)70772-8

Chen, Y., Zheng, W., Zhang, N., Gui, B., Lv, Q., Yan, J., et al. (2019). Identification of *Cryptosporidium viatorum* XVa subtype family in two wild rat species in China. *Parasitol. Vectors* 12, 502. doi: 10.1186/s13071-019-3763-6

Claerebout, E., Casaert, S., Dalemans, A. C., De Wilde, N., Levecke, B., Vercruysse, J., et al. (2009). *Giardia* and other intestinal parasites in different dog populations in Northern Belgium. *Vet. Parasitol.* 161, 41–46. doi: 10.1016/j.vetpar.2008.11.024

Damitie, M., Mekonnen, Z., Getahun, T., Santiago, D., and Leyns, L. (2018). Molecular epidemiology of *Giardia duodenalis* infection in humans in Southern Ethiopia: a triosephosphate isomerase gene-targeted analysis. *Infect. Dis. Poverty* 7, 17. doi: 10.1186/s40249-018-0397-4

de Lucio, A., Amor-Aramendía, A., Bailo, B., Saugar, J. M., Anegagrie, M., Arroyo, A., et al. (2016). Prevalence and Genetic Diversity of *Giardia duodenalis* and *Cryptosporidium* spp. among School Children in a Rural Area of the Amhara Region, North-West Ethiopia. *PLoS One* 11, e0159992. doi: 10.1371/journal.pone.0159992

Efstratiou, A., Ongerth, J. E., and Karanis, P. (2017). Waterborne transmission of protozoan parasites: Review of worldwide outbreaks-An update 2011–2016. *Water Res.* 114, 14–22. doi: 10.1016/j.watres.2017.01.036

Elwin, K., Hadfield, S. J., Robinson, G., Crouch, N. D., and Chalmers, R. M. (2012). *Cryptosporidium viatorum* n. sp. (Apicomplexa: Cryptosporidiidae) among travellers returning to Great Britain from the Indian subcontinent 2007–2011. *Int. J. Parasitol.* 42, 675–682. doi: 10.1016/j.ijpara.2012.04.016

Feleke, H., Medhin, G., Abebe, A., Beyene, B., Kloos, H., and Asrat, D. (2018). Enteric pathogens and associated risk factors among under-five children with and without diarrhea in Wegera District, Northwestern Ethiopia. *Pan. Afr. Med. J.* 29, 72. doi: 10.11604/pamj.2018.29.72.13973

Feng, Y., Ryan, U. M., and Xiao, L. (2018). Genetic Diversity and Population Structure of *Cryptosporidium*. *Trends Parasitol.* 34, 997–1011. doi: 10.1016/j.pt.2018.07.009

Feng, Y., and Xiao, L. (2011). Zoonotic potential and molecular epidemiologyof *Giardia* species and giardiasis. *Clin. Microbiol. Rev.* 24, 110–140. doi: 10.1128/CMR.00033-10

Gatei, W., Suputtamongkol, Y., Waywa, D., Ashford, R. W., Bailey, J. W., Greensill, J., et al. (2002). Zoonotic species of *Cryptosporidium* are as prevalent as the anthroponotic in HIV-infected patients in Thailand. *Ann. Trop. Med. Parasitol.* 96, 797–802. doi: 10.1179/000349802125002202

Gelanew, T., Lalle, M., Hailu, A., Pozio, E., and Cacciò, S. M. (2007). Molecular characterization of human isolates of *Giardia duodenalis* from Ethiopia. *Acta Tropica* 102, 92–99. doi: 10.1016/j.actatropica.2007.04.003

Geurden, T., Geldhof, P., Levecke, B., Martens, C., Berkvens, D., Casaert, S., et al. (2008). Mixed *Giardia duodenalis* assemblage A and E infections in calves. *Int. J. Parasitol.* 38, 259–264. doi: 10.1016/j.ijpara.2007.07.016

Guyot, K., Follet-Dumoulin, A., Lelièvre, E., Sarfati, C., Rabodonirina, M., Nevez, G., et al. (2001). Molecular characterization of *Cryptosporidium* isolates obtained from humans in France. *J. Clin. Microbiol.* 39, 3472–3480. doi: 10.1128/JCM.39.10.3472-3480.2001

Holubová, N., Zikmundová, V., Limpouchová, Z., Sak, B., Konečný, R., Hlásková, L., et al. (2019). *Cryptosporidium proventriculi* sp. n. (Apicomplexa: Cryptosporidiidae) in Psittaciformes birds. *Eur. J. Protistol.* 69, 70–87. doi: 10.1016/j.ejop.2019.03.001

Holubová, N., Tůmová, L., Sak, B., Hejzlarová, A., Konečný, R., McEvoy, J., et al. (2020). Description of *Cryptosporidium ornithophilus* n. sp. (Apicomplexa: Cryptosporidiidae) in farmed ostriches. *Parasitol. Vectors* 13, 340. doi: 10.1186/s13071-020-04191-2

Hoque, M. E., Hope, V. T., Kjellström, T., Scragg, R., and Lay-Yee, R. (2002). Risk of giardiasis in Aucklanders: a case-control study. *Int. J. Infect. Dis.* 6, 191–197. doi: 10.1016/s1201-9712(02)90110-4

Hoque, M. E., Hope, V. T., Scragg, R., and Kjellström, T. (2003). Children at risk of giardiasis in Auckland: a case-control analysis. *Epidemiol. Infect.* 131, 655–662. doi: 10.1017/s0950268803008598

Huang, D., Zhang, R., Tang, Y., Li, X., Yang, F., Wu, C., et al. (2016). Establishment and application of PCR method for detection of *Cryptosporidium. China. Tropical. Med.* 16, 8. doi: 10.13604/j.cnki.46-1064/r.2016.08.07. (in Chinese).

Hunter, P. R., Hughes, S., Woodhouse, S., Raj, N., Syed, Q., Chalmers, R. M., et al. (2004). Health sequelae of human cryptosporidiosis in immunocompetent patients. *Clin. Infect. Dis.* 39, 504–510. doi: 10.1086/422649

Hussain, G., Roychoudhury, S., Singha, B., and Paul, J. (2017). Incidence of *Cryptosporidium andersoni* in diarrheal patients from southern Assam, India: a molecular approach. *Eur. J. Clin. Microbiol. Infect. Dis.* 36, 1023–1032. doi: 10.1007/s10096-016-2887-2

Insulander, M., Silverlås, C., Lebbad, M., Karlsson, L., Mattsson, J. G., and Svenungsson, B. (2013). Molecular epidemiology and clinical manifestations of human cryptosporidiosis in Sweden. *Epidemiol. Infect.* 141, 1009–1020. doi: 10.1017/S0950268812001665

Jiang, Y., Ren, J., Yuan, Z., Liu, A., Zhao, H., Liu, H., et al. (2014). *Cryptosporidium andersoni* as a novel predominant *Cryptosporidium* species in outpatients with diarrhea in Jiangsu Province, China. *BMC Infect. Dis.* 14, 555. doi: 10.1186/s12879-014-0555-7

Johnston, A. R., Gillespie, T. R., Rwego, I. B., McLachlan, T. L., Kent, A. D., and Goldberg, T. L. (2010). Molecular epidemiology of cross-species *Giardia duodenalis* transmission in western Uganda. *PLoS Negl. Trop. Dis.* 4, e683. doi: 10.1371/journal.pntd.0000683

Karanis, P., Kourenti, C., and Smith, H. (2007). Waterborne transmission of protozoan parasites: a worldwide review of outbreaks and lessons learnt. *J. Water Health* 5, 1–38. doi: 10.2166/wh.2006.002

Khalil, S., Mirdha, B. R., Panda, A., Singh, Y., Makharia, G., and Paul, J. (2017). *Cryptosporidium* species subtypes and associated clinical manifestations in Indian patients. *Gastroenterol. Hepatol. Bed. Bench.* 10, 311–318.

Khalil, S., Mirdha, B. R., Paul, J., Panda, A., and Singh, Y. (2018). Molecular Detection and Identification of *Cryptosporidium viatorum* in a Human Immunodeficiency Virus-seropositive Patient. *J. Glob. Infect. Dis.* 10, 28–29. doi: 10.4103/jgid.jgid_26_17

Kim, M. J., Jung, B. K., Cho, J., Kim, D. G., Song, H., Lee, K. H., et al. (2016). Prevalence of Intestinal Protozoans among Schoolchildren in Suburban Areas near Yangon, Myanmar. *Korean J. Parasitol.* 54, 345–348. doi: 10.3347/kjp.2016.54.3.345

Koehler, A. V., Wang, T., Haydon, S. R., and Gasser, R. B. (2018). *Cryptosporidium viatorum* from the native Australian swamp rat *Rattus lutreolus*-An emerging zoonotic pathogen? *Int. J. Parasitol. Parasites Wildl.* 7, 18–26. doi: 10.1016/j.ijppaw.2018.01.004

Kváč, M., Vlnatá, G., Ježková, J., Horčičková, M., Konečný, R., Hlásková, L., et al. (2018). *Cryptosporidium occultus* sp. n. (Apicomplexa: Cryptosporidiidae) in rats. *Eur. J. Protistol.* 63, 96–104. doi: 10.1016/j.ejop.2018.02.001

Lalle, M., Pozio, E., Capelli, G., Bruschi, F., Crotti, D., and Cacciò, S. M. (2005). Genetic heterogeneity at the β-giardin locus among human and animal isolates of *Giardia duodenalis and* identification of potentially zoonotic subgenotypes. *Int. J. Parasitol.* 35, 207–213. doi: 10.1016/j.ijpara.2004.10.022

Leoni, F., Amar, C., Nichols, G., Pedraza-Díaz, S., and McLauchlin, J. (2006). Genetic analysis of *Cryptosporidium* from 2414 humans with diarrhoea in England between 1985 and 2000. *J. Med. Microbiol.* 55, 703–707. doi: 10.1099/jmm.0.46251-0

Levecke, B., Geldhof, P., Claerebout, E., Dorny, P., Vercammen, F., Cacciò, S. M., et al. (2009). Molecular characterisation of *Giardiaduodenalis* in captive non–human primates reveals mixed assemblage A and B infections and novel polymorphisms. *Int. J. Parasitol.* 39, 1595–1601. doi: 10.1016/j.ijpara.2009.05.013

Liu, H., Shen, Y., Yin, J., Yuan, Z., Jiang, Y., Xu, Y., et al. (2014). Prevalence and genetic characterization of *Cryptosporidium*, *Enterocytozoon*, *Giardia* and *Cyclospora* in diarrheal outpatients in China. *BMC Infect. Dis.* 14, 25. doi: 10.1186/1471-2334-14-25

Liu, A., Gong, B., Liu, X., Shen, Y., Wu, Y., Zhang, W., et al. (2020). A retrospective epidemiological analysis of human *Cryptosporidium* infection in China during the past three decades, (1987–2018). *PLoS Negl. Trop. Dis.* 14, e0008146. doi: 10.1371/journal.pntd.0008146

Morse, T. D., Nichols, R. A., Grimason, A. M., Campbell, B. M., Tembo, K. C., and Smith, H. V. (2007). Incidence of cryptosporidiosis species in paediatric patients in Malawi. *Epidemiol. Infect.* 135, 1307–1315. doi: 10.1017/S0950268806007758

Okhuysen, P. C., Chappell, C. L., Crabb, J. H., Sterling, C. R., and DuPont, H. L. (1999). Virulence of three distinct *Cryptosporidium parvum* isolates for healthy adults. *J. Infect. Dis.* 180, 1275–1281. doi: 10.1086/315033

Oliveira-Arbex, A. P., David, E. B., Oliveira-Sequeira, T. C., Bittencourt, G. N., and Guimarães, S. (2016). Genotyping of *Giardia duodenalis* isolates in asymptomatic children attending daycare centre: evidence of high risk for anthroponotic transmission. *Epidemiol. Infect.* 144, 1418–1428. doi: 10.1017/S0950268815002514

Osman, M., El Safadi, D., Cian, A., Benamrouz, S., Nourrisson, C., Poirier, P., et al. (2016). Prevalence and Risk Factors for Intestinal Protozoan Infections with *Cryptosporidium*, *Giardia*, *Blastocystis* and *Dientamoeba* among Schoolchildren in Tripoli, Lebanon. *PLoS Negl. Trop. Dis.* 10, e0004496. doi: 10.1371/journal.pntd.0004496

Plutzer, J., Lassen, B., Jokelainen, P., Djurković-Djaković, O., Kucsera, I., Dorbek-Kolin, E., et al. (2018). Review of *Cryptosporidium* and *Giardia* in the eastern part of Europe 2016. *Euro. Surveill.* 23, 16–00825. doi: 10.2807/1560-7917.ES.2018.23.4.16-00825

Prado, M. S., Cairncross, S., Strina, A., Barreto, M. L., Oliveira-Assis, A. M., and Rego, S. (2005). Asymptomatic giardiasis and growth in young children; a longitudinal study in Salvador, Brazil. *Parasitology* 131, 51–56. doi: 10.1017/s0031182005007353

Rendtorff, R. C. (1954). The experimental transmission of human intestinal protozoan parasites. II. *Giardia lamblia* cysts given in capsules. *Am. J. Hygiene* 59, 209–220. doi: 10.1093/oxfordjournals.aje.a119634

Roy, S. L., DeLong, S. M., Stenzel, S. A., Shiferaw, B., Roberts, J. M., Khalakdina, A., et al. (2004). Risk factors for sporadic cryptosporidiosis among immunocompetent persons in the United States from 1999 to 2001. *J. Clin. Microbiol.* 42, 2944–2951. doi: 10.1128/JCM.42.7.2944-2951.2004

Ryan, U. M., Bath, C., Robertson, I., Read, C., Elliot, A., McInnes, L., et al. (2005). Sheep may not be an important zoonotic reservoir for *Cryptosporidium* and *Giardia* parasites. *Appl. Environ. Microbiol.* 71, 4992–4997. doi: 10.1128/AEM.71.9.4992-4997.2005

Ryan, U., Zahedi, A., and Paparini, A. (2016). *Cryptosporidium* in humans and animals-a one health approach to prophylaxis. *Parasite Immunol.* 38, 535–547. doi: 10.1111/pim.12350

Ryan, U., Hijjawi, N., Feng, Y., and Xiao, L. (2019). *Giardia*: an under–reported foodborne parasite. *Int. J. Parasitol.* 49, 1–11. doi: 10.1016/j.ijpara.2018.07.003

Ryan, U., and Cacciò, S. M. (2013). Zoonotic potential of *Giardia*. *Int. J. Parasitol.* 43, 943–956. doi: 10.1016/j.ijpara.2013.06.001

Sánchez, A., Munoz, M., Gómez, N., Tabares, J., Segura, L., Salazar, Á, et al. (2017). Molecular Epidemiology of *Giardia*, *Blastocystis* and *Cryptosporidium* among Indigenous Children from the Colombian Amazon Basin. *Front. Microbiol.* 8, 248. doi: 10.3389/fmicb.2017.00248

Santín, M., Trout, J. M., and Fayer, R. (2007). Prevalence and molecular characterization of *Cryptosporidium* and *Giardia* species and genotypes in sheep in Maryland. *Vet. Parasitol.* 146, 17–24. doi: 10.1016/j.vetpar.2007.01.010

Savioli, L., Smith, H., and Thompson, A. (2006). *Giardia* and *Cryptosporidium* join the 'Neglected Diseases Initiative'. *Trends Parasitol.* 22, 203–208. doi: 10.1016/j.pt.2006.02.015

Scaramozzino, P., Di Cave, D., Berrilli, F., D'Orazi, C., Spaziani, A., Mazzanti, S., et al. (2009). A study of the prevalence and genotypes of *Giardia duodenalis* infecting kennelled dogs. *Vet. J.* 182, 231–234. doi: 10.1016/j.tvjl.2008.07.003

Sprong, H., Cacciò, S. M., van der Giessen, J. W.ZOOPNET network and partners (2009). Identification of zoonotic genotypes of *Giardia duodenalis*. *PLoS Negl. Trop. Dis.* 3, e558. doi: 10.1371/journal.pntd.0000558

Squire, S. A., and Ryan, U. (2017). *Cryptosporidium* and *Giardia* in Africa: current and future challenges. *Parasitol. Vectors* 10, 195. doi: 10.1186/s13071-017-2111-y

Stensvold, C. R., Elwin, K., Winiecka-Krusnell, J., Chalmers, R. M., Xiao, L., and Lebbad, M. (2015). Development and Application of a gp60-Based Typing Assay for *Cryptosporidium viatorum*. *J. Clin. Microbiol.* 53, 1891–1897. doi: 10.1128/JCM.00313-15

Su, H., Huang, X., Qin, Y., Qin, S., Hung, Z., Wu, N., et al. (2017). An investigation on the gene types of *Cryptosporidium* among HIV/AIDS patients. *J. Med. Pest Control* 33, 2. doi: 10.7629/yxdwfz201702011. (in Chinese).

Tellevik, M. G., Moyo, S. J., Blomberg, B., Hjøllo, T., Maselle, S. Y., Langeland, N., et al. (2015). Prevalence of *Cryptosporidium parvum/hominis*, *Entamoeba histolytica* and *Giardia lamblia* among Young Children with and without Diarrhea in Dar es Salaam, Tanzania. *PLoS Negl. Trop. Dis.* 9, e0004125. doi: 10.1371/journal.pntd.0004125

Traub, R. J., Monis, P. T., Robertson, I., Irwin, P., Mencke, N., and Thompson, R. C. A. (2004). Epidemiological and molecular evidence supports the zoonotic transmission of *Giardia* among humans and dogs living in the same community. *Parasitology* 128, 253–262. doi: 10.1017/s0031182003004505

Ukwah, B. N., Ezeonu, I. M., Ezeonu, C. T., Roellig, D., and Xiao, L. (2017). *Cryptosporidium* species and subtypes in diarrheal children and HIV-infected persons in Ebonyi and Nsukka, Nigeria. *J. Infect. Dev. Ctries.* 11, 173–179. doi: 10.3855/jidc.8034

Waldron, L. S., Dimeski, B., Beggs, P. J., Ferrari, B. C., and Power, M. L. (2011). Molecular epidemiology, spatiotemporal analysis, and ecology of sporadic human cryptosporidiosis in Australia. *Appl. Environ. Microbiol.* 77, 7757–7765. doi: 10.1128/AEM.00615-11

Wang, K., Gazizova, A., Wang, Y., Zhang, K., Zhang, Y., Chang, Y., et al. (2019). First Detection of *Cryptosporidium* spp. in Migratory Whooper Swans (*Cygnus cygnus*) in China. *Microorganisms* 8, 6. doi: 10.3390/microorganisms8010006

Wang, Y., Gonzalez-Moreno, O., Roellig, D. M., Oliver, L., Huguet, J., Guo, Y., et al. (2019). Epidemiological distribution of genotypes of *Giardia duodenalis* in humans in Spain. *Parasitol. Vectors* 12, 432. doi: 10.1186/s13071-019-3692-4

Xiao, L., and Fayer, R. (2008). Molecular characterisation of species and genotypes of *Cryptosporidium* and *Giardia* and assessment of zoonotic transmission. *Int. J. Parasitol.* 38, 1239–1255. doi: 10.1016/j.ijpara.2008.03.006

Xiao, L., and Feng, Y. (2017). Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*. *Food. Waterborne Parasitol.* 8–9, 14–32. doi: 10.1016/j.fawpar.2017.09.002

Xu, N., Liu, H., Jiang, Y., Yin, J., Yuan, Z., Shen, Y., et al. (2020). First report of *Cryptosporidium viatorum* and *Cryptosporidium occultus in* humans in China, and of the unique novel *C. viatorum* subtype XVaA3h. *BMC Infect. Dis.* 20, 16. doi: 10.1186/s12879-019-4693-9

Yu, F., Li, D., Chang, Y., Wu, Y., Guo, Z., Jia, L., et al. (2019). Molecular characterization of three intestinal protozoans in hospitalized children with different disease backgrounds in Zhengzhou, central China. *Parasitol. Vectors* 12, 543. doi: 10.1186/s13071-019-3800-5

Zhao, Z., Dong, H., Wang, R., Zhao, W., Chen, G., Li, S., et al. (2014). Genotyping and subtyping *Cryptosporidium parvum* and *Giardia duodenalis* carried by flies on dairy farms in Henan, China. *Parasitol. Vectors* 7, 190. doi: 10.1186/1756-3305-7-190

Zhao, W., Zhou, H., Huang, Y., Xu, L., Rao, L., Wang, S., et al. (2019). *Cryptosporidium* spp. in wild rats (*Rattus* spp.) from the Hainan Province, China: Molecular detection, species/genotype identification and implications for public health. *Int. J. Parasitol. Parasites Wildl.* 9, 317–321. doi: 10.1016/j.ijppaw.2019.03.017

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.

Copyright © 2020 Wu, Gong, Liu, Jiang, Cao, Yao, Li, Liu and Shen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 30 November 2020
doi: 10.3389/fmicb.2020.584717

Comparative Genomic Analyses Reveal a Specific Mutation Pattern Between Human Coronavirus SARS-CoV-2 and Bat-CoV RaTG13

Longxian Lv[1†], Gaolei Li[2†], Jinhui Chen[3], Xinle Liang[2] and Yudong Li[2]*

[1] State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China, [2] Department of Biological Engineering, School of Food Science and Biotechnology, Zhejiang Gongshang Univeristy, Hangzhou, China, [3] College of Biological and Chemical Engineering, Jiaxing University, Jiaxing, China

Background: The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China, rapidly grew into a global pandemic. How SARS-CoV-2 evolved remains unclear.

Methods: We performed a comprehensive analysis using the available genomes of SARS-CoV-2 and its closely related coronaviruses.

Results: The ratio of nucleotide substitutions to amino acid substitutions of the spike gene (9.07) between SARS-CoV-2 WIV04 and Bat-CoV RaTG13 was markedly higher than that between other coronaviruses (range, 1.29–4.81); the ratio of non-synonymous to synonymous substitution rates (dN/dS) between SARS-CoV-2 WIV04 and Bat-CoV RaTG13 was the lowest among all the performed comparisons, suggesting evolution under stringent selective pressure. Notably, the relative proportion of the T:C transition was markedly higher between SARS-CoV-2 WIV04 and Bat-CoV RaTG13 than between other compared coronaviruses. Codon usage is similar across these coronaviruses and is unlikely to explain the increased number of synonymous mutations. Moreover, some sites of the spike protein might be subjected to positive selection.

Conclusions: Our results showed an increased proportion of synonymous substitutions and the T:C transition between SARS-CoV-2 and RaTG13. Further investigation of the mutation pattern mechanism would contribute to understanding viral pathogenicity and its adaptation to hosts.

Keywords: SARS-CoV-2, synonymous mutation, codon usage, natural selection, Bat-CoV

OPEN ACCESS

Edited by:
Jianying Gu,
College of Staten Island,
United States

Reviewed by:
Zhao Zhang,
University of Texas Health Science Center at Houston, United States
Kaixiong Ye,
University of Georgia, United States

****Correspondence:***
Yudong Li
lyd@zjsu.edu.cn

[†]These authors share first authorship

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

Received: *17 July 2020*
Accepted: *05 November 2020*
Published: *30 November 2020*

Citation:
Lv L, Li G, Chen J, Liang X and Li Y (2020) Comparative Genomic Analyses Reveal a Specific Mutation Pattern Between Human Coronavirus SARS-CoV-2 and Bat-CoV RaTG13. Front. Microbiol. 11:584717. doi: 10.3389/fmicb.2020.584717

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also known as 2019-nCoV, is a novel coronavirus (CoV) isolated from patients with pneumonia in China in 2019. SARS-CoV-2 has a similar incubation period (median, 3 days) and a relatively lower fatality when compared with SARS-CoV or MERS-CoV (Jiang et al., 2020), but the reproductive number of SARS-CoV-2 is estimated to be higher than that of SARS-CoV (Liu et al., 2020). Moreover, some laboratory-confirmed symptomatic cases lack apparent cough, fever, or radiologic manifestations, making it

difficult to identify all infected patients in a timely and accurate manner (Xu et al., 2020). As of July 17, 2020, patients infected by SARS-CoV-2 have been diagnosed in more than 200 countries, and more than 13 million confirmed cases and 560,000 deaths associated with SARS-CoV-2 infection have been reported worldwide.

The genetic information of a virus is essential for its classification and traceability and its pathogenicity. At the whole genome level, the sequence identity of SARS-CoV-2 is 50% to that of MERS-CoV; 79% to that of SARS-CoV; 88% to those of two bat-derived SARS-like coronaviruses, Bat-SL-CoVZC45 and Bat-SL-CoVZXC21 (collected in 2018 in Zhoushan, China); and 96% to that of Bat-CoV RaTG13 (collected in 2013 in Yunnan, China) (Zhang L. et al., 2020; Zhou et al., 2020). Each genome of all SARS-CoV-2 strains now submitted online contains nearly 29,900 nucleotides (nt), which are predicted with at least 14 open reading frames (ORFs) (5′–3′), such as *ORF1ab* (*P*, 21,291 nt), *spike* (*S*, 3,822 nt), *ORF3a* (828 nt), *envelope* (*E*, 228 nt), *membrane* (*M*, 669 nt), *ORF8* (366 nt), and *nucleocapsid* (*N*, 1,260 nt) (Wu et al., 2020). Among them, the spike gene encodes a glycoprotein that is crucial to determine host tropism and transmission capacity and is highly divergent compared with that of Bat-CoV RaTG13 (93.1% nucleotide identity) (Lu et al., 2020; Wu et al., 2020).

Generally, the rates of nucleotide substitution of RNA viruses are faster than those of their hosts, and this rapid evolution is mainly shaped by natural selection (mostly purifying selection) (Lin et al., 2019). Genetic mutations such as nucleotide substitutions, deletions, and insertions have been frequently reported when comparing SARS-CoV-2 with other viruses (Lu et al., 2020; Wu et al., 2020; Zhou et al., 2020). In this study, we investigated the potential mutation pattern of SARS-CoV-2 by comprehensive comparative genomic analysis of non-synonymous/synonymous substitutions, relative synonymous codon usage (RSCU), and selective pressure to explore their potential roles in virus evolution.

MATERIALS AND METHODS

Sequence Data

The SARS-CoV-2 reference genomes Wuhan-Hu-1 (NC_045512) and WIV04 (MN996528) were downloaded from the GenBank database. Twenty-one closely related coronavirus complete genome sequences and their coding sequences were also downloaded from the GenBank database (**Table 1**).

Phylogenetic Analysis

Genome sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004), followed by manual adjustment using BioEdit v7.2.5. Phylogenetic analyses of the complete genome were performed using the maximum-likelihood method and general time-reversible model of nucleotide substitution with gamma-distributed rates among sites (GTR + G) in RAxML v8.1.21 (Stamatakis, 2014). Support for the inferred relationships was evaluated using bootstrap analysis with 1,000 replicates, and trees were rooted using the alpha-coronavirus lineage as an outgroup.

The coding sequences were translated and aligned using the MEGA X program (Kumar et al., 2018), and then codon-based sequence alignment was used for further analysis. Phylogenetic analyses of coding sequences were performed using the MEGA X software. The changes in amino acids or nucleotides for each coding sequence were analyzed using in-house Perl scripts. Both NT and AA changes were counted by comparing to the reference strain in each CoV lineage.

Estimation of Synonymous and Non-synonymous Substitution Rates

The number of synonymous substitutions per synonymous site (dS) and the number of non-synonymous substitutions per non-synonymous site (dN) for each coding region were calculated using the Nei–Gojobori method (Jukes–Cantor) in the PAML package.

The adaptive evolution server[1] was used to identify the eventual sites of positive selection. For this purpose, the mixed-effects model of evolution (MEME), which allows the distribution of dN/dS (ω) to vary from site to site and from branch to branch at a site, was used (Murrell et al., 2012). This test allowed us to infer episodic and pervasive positive selection at individual sites.

Synonymous Codon Usage Analysis

To investigate the potential RSCU bias of the spike protein from SARS-CoV-2 and its closely related coronaviruses, the coding sequence of the spike protein in these coronaviruses was calculated with CodonW 1.4.4[2]. The RSCU of human genes

[1]http://www.datamonkey.org/

[2]http://codonw.sourceforge.net/

TABLE 1 | Coronavirus genome sequences used in this study.

Strain name	Accession number	Host
WIV04	MN996528	Human
SNU01	MT039890	Human
RaTG13	MN996532	Bat
CoVZC45	MG772933	Bat
Tor2	AY274119	Human
civet007	AY572034	Civet
WIV1	KF367457	Bat
LYRa11	KF569996	Bat
HKU3-8	GQ153543	Bat
BtKY72	KY352407	Bat
Zhejiang2013	NC_025217	Bat
EMC2012	JX869059	Human
CAMEL-363	KJ713298	Camel
RSA2011	KC869678	Bat
bCoV-ENT	NC_003045	Bovine
MHV-A59	NC_001846	Mouse
HKU1	NC_006577	Bat
FarmA	MF094681	Pig
HKU2	NC_009988	Bat
229E	NC_002645	Human
PEDV	NC_003436	Pig

was retrieved from the Codon Usage Database[3]. The potential relationships among these sequences were calculated using a squared Euclidean distance [$d_{ik} = \sum_{j=1}^{p}(X_{ij}^2 - X_{kj}^2)$]. In addition to RSCU, the effective number of codons (ENc) was used as a simple metric to verify codon bias and explore the source of the virus.

Statistical Analysis

Statistical analyses were performed using the R statistical package (version 3.2.2). Chi-squared test was used to compare any two data sets, and the data were considered significantly different if the two-tailed p value was less than 0.05.

RESULTS

The Mutation Pattern Between SARS-CoV-2 and Its Closely Related Coronaviruses

From December 2019 to February 2020, the genome sequences of 108 strains of SARS-CoV-2 virus were submitted to the global initiative on sharing avian influenza database (GISAID) worldwide. Compared with the standard SARS-CoV-2 strain WIV04, 98 point mutations were detected at 93 nucleotide sites in all SARS-CoV-2 strains with genome sequences available on February 25, 2020. However, only 58 of these nucleotide mutations caused changes in amino acids. Among them, 15 nucleotide substitutions at 14 sites caused changes in 7 amino acids of the spike protein.

The newly identified SARS-CoV-2 strain WIV04 genome sequence is closely related to Bat-CoV RaTG13 and Bat-SL-CoVZC45, which were collected from the horseshoe bat *Rhinolophus affinis* (Zhou et al., 2020). Compared with the RaTG13 genome, many nucleotide substitutions are observed, but only five small insertion and deletion (indel) mutations. The largest insert segment in the WIV04 genome was the "CGGCGGGCACGT" sequence, which is located near the boundary of the S1 and S2 regions of the spike protein. Interestingly, only synonymous mutations are observed near this insertion sequence (**Figure 1**). Compared with the Bat-SL-CoVZC45 genome, this insert segment is detected in the SARS-CoV-2 genome as well, but non-synonymous mutations are also observed around this insert sequence.

Next, we compared the proportion of synonymous mutations in the spike gene between WIV04 and RaTG13 or Bat-SL-CoVZC45. The ratio of nucleotide substitutions (263 NT) to amino acid substitutions (29 AA) was 9.07 from WIV04 to RaTG13, significantly higher than the ratio (3.91, 864/221) from WIV04 to Bat-SL-CoVZC45 ($p < 0.05$) (**Figure 2**). Similar results were observed in comparisons using their whole genomes. Consequently, the proportion of synonymous mutations (∼40%) among all currently reported SARS-CoV-2 strains is similar to that between WIV04 and Bat-SL-CoVZC45 (39.1%) but is dramatically lower than that between WIV04 and RaTG13 (90.7%).

Furthermore, the proportion of T-to-C (T:C) transitions in the whole nucleotide mutation was markedly higher between WIV04 and RaTG13 than between others (**Figure 3**). Interestingly, CoVs lacking a 3′-to-5′ exoribonuclease (ExoN) accumulate 15- to 20-fold more A:G and U:C transitions (Smith et al., 2013). Thus, the specific T:C transition pattern between SARS-CoV-2 and RaTG13 might have resulted from the loss function of ExoN in the process of evolution.

The Mutation Pattern Between SARS-CoV-2 and RaTG13 Is Unique Across Coronavirus Species

To investigate whether the increased ratio of the nucleotide to amino acid changes between SARS-CoV-2 and RaTG13 is unique across all coronavirus species, we further compared their alterations in nucleotides and amino acids with other representative coronaviruses. Phylogenetic analysis of SARS-CoV-2 and its 20 closely related coronaviruses formed four well-supported clades (**Figure 2**). The two SARS-CoV-2 strains WIV04 (from a Wuhan patient) and SNU01 (from a Korean patient) were clustered with SARS-CoV-related strains to form clade 1, belonging to beta-coronavirus lineage B. MERS-CoV (EMC2012) from humans, CAMEL-363 from camels, and RSA2011 from bats formed clade 2, belonging to beta-coronavirus lineage C. Beta-coronavirus lineage A is formed by the representative bovine coronavirus (bCoV-ENT), murine hepatitis virus (MHV-A59), and human coronavirus HKU1. The last clade comprised four representative alpha-coronaviruses. Notably, the sequence identities between closely related representative viruses, such as WIV04 and RaTG13, Tor2 and WIV1, and SADS FarmA and HKU1, were nearly 95%.

Next, we compared the alterations in the nucleotides and amino acids between these coronaviruses. As described previously, the ratio of the number of nucleotide changes to amino acid changes (NT/AA) was as high as 9.07 when SARS-CoV-2 was compared with Bat-CoV RaTG13 (**Figure 2**). However, this ratio was less than 5.0 when comparing other human coronaviruses with their similar animal coronaviruses. For example, the ratio of human SARS-CoV Tor2 to bat SARS-like CoV LYRa11 was 4.81, the ratio of human MERS-CoV EMC2012 to bat SARS-like CoV RSA2011 was 2.80, and the ratio of human coronavirus 229E to bat coronavirus HKU2 was 4.55. These results indicate that the relative level of synonymous substitutions between human SARS-CoV-2 and its possible animal origin (RaTG13) is much higher than that between other human coronaviruses and their related animal strains.

Codon Usage Is Similar Across Beta-Coronavirus Lineages

Different organisms, including different protein-coding genes of the same species, have different frequencies of codon usage (Zhou et al., 2020). The RSCU bias of coronaviruses will reveal the difference in their host source. We calculated the distance of RSCU between the spike genes of 20 representative coronaviruses

[3]http://www.kazusa.or.jp/codon/

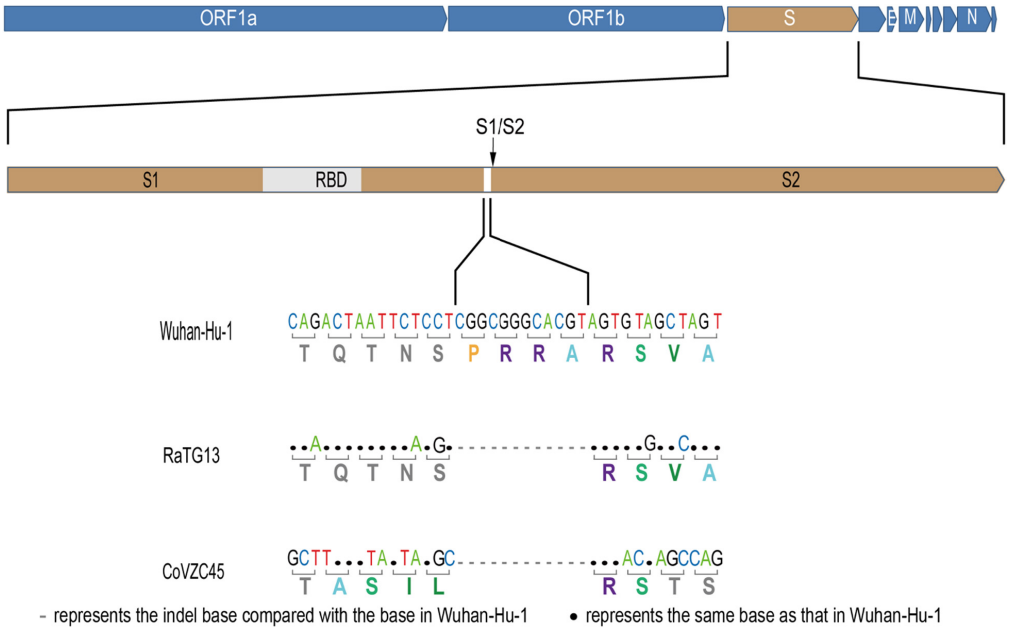


FIGURE 1 | Base mutations of the largest insertion in the spike sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) against the most closely related SARS-like CoV RaTG13 and Bat-SL-CoVZC45.

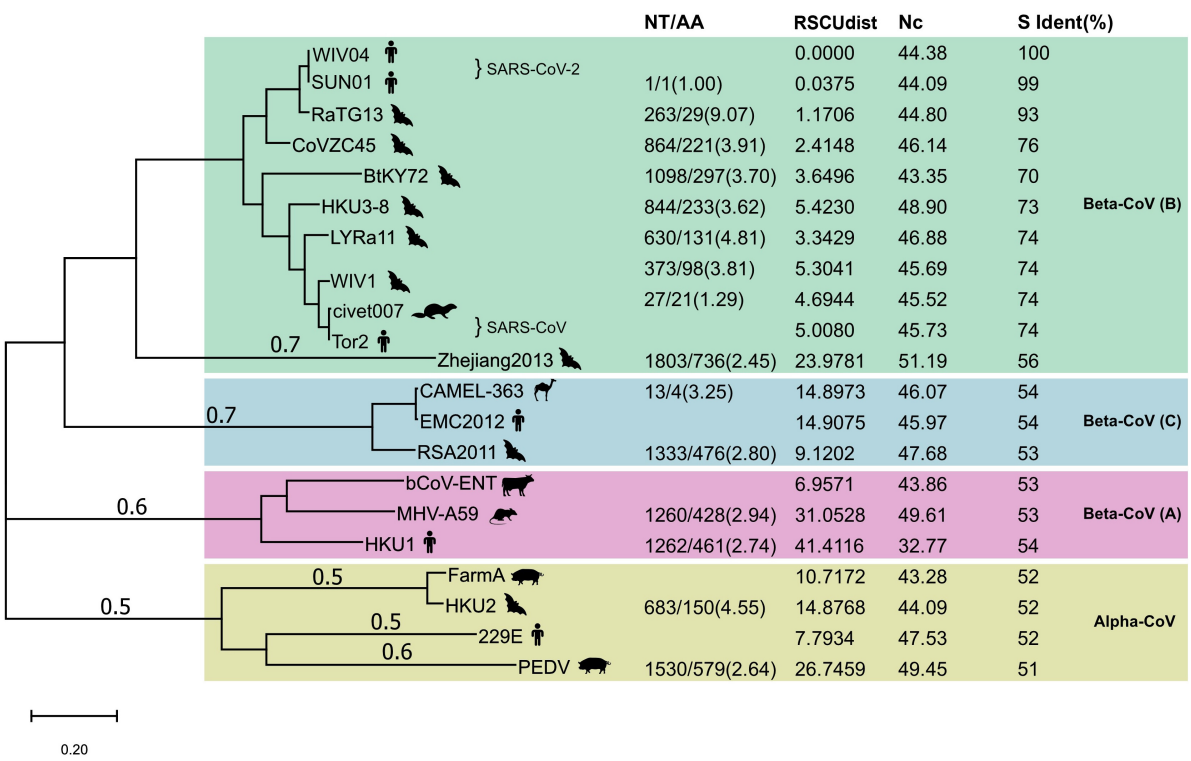


FIGURE 2 | Maximum-likelihood phylogenetic tree of 21 coronavirus strains. The tree was built with the full genome and rooted against the alpha-CoV clade. NT, number of nucleotide changes; AA, number of amino acid changes. Both NT and AA were the changes from the reference strain that had no numerical value in the NT/AA column in each CoV lineage. RSCUdist, RSCU distance of the spike protein; ENc, effective number of codons; S ident (%), percent identity of the S protein sequence between different coronaviruses. The NT/AA ratio between WIV04/RaTG13 was significantly higher (chi-squared test; $p < 0.01$) than that of other strain pairs.

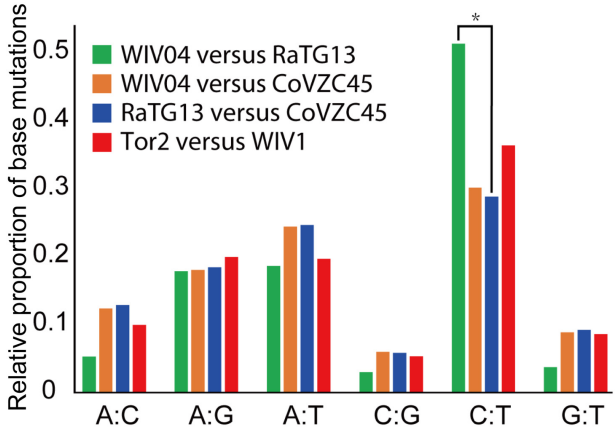


FIGURE 3 | Frequency of point mutations observed in the spike gene. The frequency is the number of point mutations in each category (A:C, A:G, A:T, C:G, C:T, and G:T) divided by the total mutations. "A:C" indicates that nucleotide A was changed to C, or C changed to A. *$p < 0.05$.

and SARS-CoV-2. The codon usage difference (distance of RSCU) between SARS-CoV-2 WIV04 and Bat-CoV RaTG13 was 1.17, which was the lowest except for SNU01 (another strain of SARS-CoV-2), indicating that their codon preference was almost the same; the second lowest codon usage difference was 2.41, which was detected between SARS-CoV-2 WIV04 and Bat-SL-CoVZC45 (**Figure 2**). The coronaviruses in the same beta-CoV lineage (B) have a relatively close distance, except CoV Zhejiang2013. The codon usage difference of the spike gene between human coronavirus HKU1 and bovine bCoV-ENT sequence had the largest difference, with a ratio of 41.41.

Codon usage bias in a gene can be effectively measured by determining the ENc. The lower ENc values represent high codon bias with low numbers of synonymous codons used for the amino acids, and a gene with strong codon usage bias may have an ENc value less than 35. The ENc value of the WIV04 spike gene was 44.38 (**Figure 2**), which is similar to those of RaTG13 and other bat coronaviruses in the B, indicating that the high synonymous mutation was unlikely to be determined by codon usage bias.

The Nucleotide Substitutions Between SARS-CoV-2 and RaTG13 Are Affected by Stronger Purifying Selection

To infer whether the retention of mutations is supported or hindered by natural selection, we further studied the non-synonymous substitution rate (dN) and synonymous substitution rate (dS) in the spike gene (**Table 2**). Generally, positive (Darwinian) selection increases, but negative (purifying) selection decreases the ratios of non-synonymous to synonymous substitution rates (dN/dS). Our results showed that both dN and dS of the *S* gene of SARS-CoV-2 WIV04 versus Bat-SARr-CoV RaTG13 were the lowest among all typical coronaviruses, while those of SARS-CoV Tor2 versus bat SARS-like coronavirus WIV1 were the second lowest. When the ratio of dN to dS of the spike gene is compared, all the tested dN/dS values are less than 1, indicating that these non-synonymous mutations are harmful, and negative selection will reduce their retention speed (**Table 2**). Among them, the dN/dS of the spike gene of SARS-CoV-2 WIV04 versus that of Bat-SARr-CoV RaTG13 was 0.04, which was the lowest among all comparisons, reconfirming that the rate of synonymous mutation was extensively high between WIV04 and RaTG13 strains. Moreover, the dN/dS rates of the polyprotein (ORF1ab) and nucleocapsid (N) genes were similar to that of the spike gene (**Supplementary Table 1**).

Because the receptor-binding domain (RBD) of the spike protein is involved in interacting with human angiotensin-converting enzyme 2 (ACE2) protein, the RBD region is thought to be a preferential target of natural selection (Forni et al., 2016). Consistent with this hypothesis, our results showed that both dN and dS of the RBD region are higher than those of the whole spike gene region across all the virus pairs used in this study (**Table 2**). Notably, the dN/dS ratio of the RBD region in SARS-CoV-2 WIV04 was dramatically increased by approximately threefold compared with the full spike region. Consequently, these mutations might be subjected to Darwinian selection or relaxation of purifying selection.

Furthermore, MEME analysis was performed to detect positive selection on the spike gene. Significant ($p < 0.05$) pervasive episodic selection was detected in three sites (48th, 254th, and 330th positions using the reference sequence of WIV04) on the common ancestor of WIV04 and RaTG13

TABLE 2 | Comparison of the evolutionary rate of the spike protein and its RBD region between different coronavirus strains.

Coronavirus strains for pair comparison		Spike			Receptor-binding domain (RBD)		
		dN/Ds	dN	dS	dN/dS	dN	Ds
WIV04	RaTG13*	0.04	0.014	0.31	0.1165	0.064	0.5494
WIV04	CoVZC45	0.11	0.13	1.19	0.1067	0.2213	2.0751
RaTG13	CoVZC45	0.12	0.13	1.08	0.1067	0.2213	2.0751
WIV04	Tor2	0.11	0.17	1.50	0.095	0.028	0.2945
Tor2	WIV1*	0.17	0.05	0.32	0.1592	0.1997	1.2545
FarmA	HKU2*	0.09	0.097	1.08	0.0835	0.1114	1.3342
EMC2012	RSA2011	0.21	0.29	1.38	0.4015	0.6464	1.61

The whole genome sequence identity between these paired strains was larger than 95%.

lineages. At the 254th position of the spike amino acid sequence, a histidine residue is present instead of a phenylalanine residue; at the 330th amino acidic position in the WIV04 sequence, a glutamine residue is present instead of a valine residue. The results described above support the action of positive selection on some sites of the spike gene during the recent evolution of SARS-CoV-2 and RaTG13.

DISCUSSION

In this study, we first observed that the proportion of synonymous substitutions was similar to that of non-synonymous substitutions within currently available SARS-CoV-2 strains. According to the random drift hypothesis (Castellano et al., 2018), these nucleotide differences among different SARS-CoV-2 strains may primarily result from neutral evolution. In short, no powerful factor exists to force SARS-CoV-2 to evolve in a certain direction. However, strict precautions should be taken against the strong factors that may cause directional variation of SARS-CoV-2 both in the natural environment and during infection treatment.

Second, our results showed that synonymous mutations are dramatically elevated between SARS-CoV-2 and RaTG13. The relative proportion of synonymous substitutions between human SARS-CoV-2 and its possible animal origin (RaTG13) is much higher than that between other human coronaviruses and their potential animal sources. These results indicated that the SARS-CoV-2 strains might undergo stronger purifying selection after diverging from their common ancestor. Interestingly, the nucleotide mutations were enriched in the T:C transition. The specific mutation pattern may be caused by the inactivation of RNA 3′-to-5′ ExoN (Smith et al., 2013). This increased T:C mutation implies that the ExoN of SARS-CoV-2 may be deactivated compared with that of RaTG13. Moreover, RNA mutagen 5-fluorouracil (5-FU) treatment can increase U:C and A:G transitions as well (Smith et al., 2013). Therefore, the underlying mechanisms of such potential mutations between SARS-CoV-2 and RaTG13 require further investigation in the future.

Previous studies on codon usage bias between viruses and their hosts have suggested that viruses tend to evolve codon usage bias comparable to their hosts (Bahir et al., 2009). Generally, RNA viruses usually comprise high codon usage bias, which helps in replication and host adaption with preferred codons. However, our results showed that codon usage was similar between SARS-CoV-2 and other strains in beta-coronavirus lineage B, suggesting that the presumptive mutation pattern was not determined by codon usage bias.

The non-synonymous (dN) to synonymous substitutions (dS) ratio in protein-coding genes is commonly used to detect the selection pressure during gene evolution. A dN/dS ratio larger than 1 indicates positive selection, while a dN/dS ratio less than 1 indicates negative selection acting on protein-coding genes. Our results showed that the dN/dS between SARS-CoV-2 and RaTG13 was less than 0.1 and significantly lower than that of other paired strain comparisons. These results indicate that the SARS-CoV-2 virus exhibits extraordinarily stringent negative selection pressure if it evolved from RaTG13. By contrast, the relatively high dN/dS ratio in the RBD region of the spike protein suggests that the selective pressure acting on this region is relaxed, and some sites may be undergoing positive selection. This increased evolutionary rate can be explained by the important function of the spike protein, which participates in host-specific recognition and undergoes several drastic changes during virus infection. For example, its large parts are cleaved during infection by cellular proteases, exposing the receptors to activate viral attachment to the host (Lu et al., 2015). However, this result should be addressed carefully because the RBD region of the spike gene from SARS-CoV-2 is divergent from that of RaTG13, suggesting that it may have originated from homologous recombination between RaTG13 and one yet-unknown coronavirus (Zhang C. et al., 2020).

In summary, through comprehensive comparative analysis between SARS-CoV-2 and other coronaviruses, we found that synonymous mutations were dramatically elevated between SARS-CoV-2 and RaTG13 compared with other coronavirus strains, and nucleotide mutations were enriched in the T:C transition. Because SARS-CoV-2 is supposed to originate from Bat-CoV RaTG13, the increased synonymous substitution between SARS-CoV-2 and the RaTG13 strain suggests that the SARS-CoV-2 genome should be under stringent negative (purifying) selection. Moreover, the mechanism underpinning the increased T:C mutations requires further investigation.

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/s.

AUTHOR CONTRIBUTIONS

YL conceived and designed the study. JC and XL completed the data collection. LL and GL conducted statistics and analysis of the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 31671836 and 81570512) and the National Key Research and Development Program of China (grant number 2018YFC2000500).

ACKNOWLEDGMENTS

We thank all the laboratories submitting the genome sequences of SARS-CoV-2 to GISAID or GenBank database for public

research. YL would like to thank Wen-hsiung Li for hosting him as a visiting student to learn molecular evolution at his laboratory at the University of Chicago. This manuscript has been released as a preprint at bioRxiv (Lv et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.584717/full#supplementary-material

REFERENCES

Bahir, I., Menachem, F., Yosef, P., and Michal, L. (2009). Viral adaptation to host: a proteome–based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* 5:311. doi: 10.1038/msb.2009.71

Castellano, D., Jennifer, J., and Adam, E. (2018). Nearly neutral evolution across the drosophila melanogaster genome. *Mol. Biol. Evol.* 35, 2685–2694. doi: 10.1093/molbev/msy164

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Forni, D., Rachele, C., Alessandra, M., Uberto, P., Nasser, A., Mario, C., et al. (2016). Extensive positive selection drives the evolution of nonstructural proteins in lineage C betacoronaviruses. *J. Virol.* 90, 3627–3639. doi: 10.1128/JVI.02988-15

Jiang, X., Simon, R., and Luo, M. (2020). Does SARS-CoV-2 has a longer incubation period than SARS and MERS? *J. Med. Virol.* 92, 476–478. doi: 10.1002/jmv.25708

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

Lin, J., Bhattacharjee, M. J., Yu, C., Tseng, Y. Y., and Li, W. (2019). Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19009–19018. doi: 10.1073/pnas.1907626116

Liu, Y., Gayle, A. A., Wilder-Smith, A., and Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* 27:taaa021. doi: 10.1093/jtm/taaa021

Lu, G., Wang, Q., and Gao, G. F. (2015). Bat-to-human: spike features determining 'Host Jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol.* 23, 468–478. doi: 10.1016/j.tim.2015.06.003

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, 566–574. doi: 10.1016/S0140-6736(20)30251-8

Lv, L., Li, G., Chen, J., Liang, X., and Li, Y. (2020). Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. *bioRxiv[Preprint]* doi: 10.1101/2020.02.27.969006

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764. doi: 10.1371/journal.pgen.1002764

Smith, E. C., Blanc, H., Surdel, M. C., Vignuzzi, M., and Denison, M. R. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. *PLoS Pathog.* 9:e1003565. doi: 10.1371/journal.ppat.1003565

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

Wu, F., Zhao, S., Yu, B., Chen, Y., Wang, W., Song, Z., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3

Xu, X., Wu, X., Jiang, X., Xu, K., Ying, L., Ma, C., et al. (2020). Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series. *BMJ* 368:m606. doi: 10.1136/bmj.m606

Zhang, C., Zheng, W., Huang, X., Bell, E. W., Zhou, X., and Zhang, Y. (2020a). Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *J. Proteome Res.* 19, 1351–1360. doi: 10.1021/acs.jproteome.0c00129

Zhang, L., Shen, F., Chen, F., and Lin, Z. (2020b). Origin and evolution of the 2019 novel coronavirus. *Clin. Infect. Dis.* 71, 882–883. doi: 10.1093/cid/ciaa112

Zhou, P., Yang, X., Wang, X., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

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.

Copyright © 2020 Lv, Li, Chen, Liang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 08 January 2021
doi: 10.3389/fcimb.2020.610985

Genomic Epidemiology of Antimalarial Drug Resistance in *Plasmodium falciparum* in Southern China

Fang Huang[1], Christopher G. Jacob[2], Shannon Takala-Harrison[3], Matthew Adams[3], Heng-Lin Yang[4], Hui Liu[4], Zhi-Gui Xia[1], Shui-Sen Zhou[1], Lin-Hua Tang[1] and Christopher V. Plowe[5]**

[1] National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Chinese Center for Tropical Diseases Research, Key Laboratory of Parasite and Vector Biology, Ministry of Health, WHO Collaborating Centre for Tropical Diseases, National Center for International Research on Tropical Diseases, Shanghai, China, [2] Malaria Programme, Wellcome Sanger Institute, Cambridge, United Kingdom, [3] Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, MD, United States, [4] Malaria Department, Yunnan Institute of Parasitic Diseases, Puer, China, [5] Duke Global Health Institute, Duke University, Durham, NC, United States

OPEN ACCESS

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

Reviewed by:
Guoding Zhu,
Jiangsu Institute of Parasitic Diseases (JIPD), China
Jun Li,
Florida International University, United States

***Correspondence:**
Fang Huang
huangfang@nipd.chinacdc.cn
Christopher V. Plowe
plowe.chris@gmail.com

Specialty section:
This article was submitted to Parasite and Host, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: 28 September 2020
Accepted: 20 November 2020
Published: 08 January 2021

Citation:
Huang F, Jacob CG, Takala-Harrison S, Adams M, Yang H-L, Liu H, Xia Z-G, Zhou S-S, Tang L-H and Plowe CV (2021) Genomic Epidemiology of Antimalarial Drug Resistance in Plasmodium falciparum in Southern China. Front. Cell. Infect. Microbiol. 10:610985. doi: 10.3389/fcimb.2020.610985

Emerging artemisinin resistance in Southeast Asia poses a significant risk to malaria control and eradication goals, including China's plan to eliminate malaria nationwide by 2020. *Plasmodium falciparum* was endemic in China, especially in Southern China. Parasites from this region have shown decreased susceptibility to artemisinin and delayed parasite clearance after artemisinin treatment. Understanding the genetic basis of artemisinin resistance and identifying specific genetic loci associated with this phenotype is crucial for surveillance and containment of resistance. In this study, parasites were collected from clinical patients from Yunnan province and Hainan island. The parasites were genotyped using a *P. falciparum*-specific single nucleotide polymorphism (SNP) microarray. The SNP profiles examined included a total of 27 validated and candidate molecular markers of drug resistance. The structure of the parasite population was evaluated by principal component analysis by using the EIGENSOFT program, and ADMIXTURE was used to calculate maximum likelihood estimates for the substructure analysis. Parasites showed a high prevalence of resistance haplotypes of pfdhfr and pfdhps and moderate prevalence of pfcrt. There was no mutation identified on pfmdr1. Candidate SNPs on chromosomes 10, 13, and 14 that were associated with delayed parasite clearance showed a low prevalence of mutants. Parasites from Southern China were clustered and separated from those from Southeast Asia. Parasites from Yunnan province were substructured from parasites from Hainan island. This study provides evidence for a genomic population with drug resistance in Southern China and also illustrates the utility of SNP microarrays for large-scale parasite molecular epidemiology.

Keywords: *Plasmodium falciparum*, antimalarial drugs, artemisinin resistance, microarrays, Southern China

INTRODUCTION

Malaria remains one of the most serious infectious diseases in the world. According to the world malaria report, an estimated 228 million cases of malaria worldwide were reported in 2018 compared with 251 million in 2010 and an estimated 405,000 deaths globally in 2018 compared with 585,000 in 2010 (WHO, 2019). *Plasmodium falciparum* has developed resistance to almost all the antimalarial drugs that have been used. More recently, multidrug resistance, including to artemisinin derivatives and partner drugs of *P. falciparum*, has emerged and spread in Southeast Asia (Ashley et al., 2014; Takala-Harrison et al., 2015; Srimuang et al., 2016). The countries of the Greater Mekong Subregion (GMS) are pursuing malaria elimination with an aim to achieve regional malaria elimination by 2030 driven by this emerging multidrug-resistant *P. falciparum* (WHO, 2015). China has launched the National Malaria Elimination Action Plan 2010–2020 with an ultimate goal to interrupt local malaria transmission by 2020. Southern China was historically the main *P. falciparum* endemic region in China, especially Yunnan and Hainan Provinces. Yunnan Province, which shares borders with Myanmar, Vietnam, and Laos, is the key focus of the national malaria elimination program. China was the first country to use artemisinin, and its wide-scale use began in the early 1990s (Tu, 2011). The national malaria treatment policy of China was updated in 2016, and artemisinin-based combinations are first-line drugs used to treat *P. falciparum* malaria, including dihydroartemisinin-piperaquine, artesunate-amodiaquine, and artemisinin-piperaquine (NHC, 2016). However, parasites in Southern China have shown decreased *in vitro* susceptibility to artemisinin by the ring stage assay and delayed parasite clearance after artemisinin treatment (Huang et al., 2012b; Huang et al., 2015; Wang et al., 2015).

Several molecular markers of *P. falciparum* resistance have been identified. Mutations in *pfcrt*, which encodes a protein located on the digestive vacuole membrane, are responsible for chloroquine (CQ) resistance or treatment failure (Picot et al., 2009; Lakshmanan et al., 2014), and the K76T allele in *pfcrt* has been used for the surveillance of clinical CQ resistance (Djimde et al., 2001; Picot et al., 2009). The *pfmdr1* gene encodes the plasmodial homologue of mammalian multidrug resistance transporters linked with *antimalarial drug resistance (*Holmgren et al., 2007*;* Dahlstrom et al., 2009*;* Vinayak et al., 2010*;* Ferreira et al., 2011*). The single nucleotide polymorphisms (SNPs) at codons N86Y, Y184F, S1034C, N1042D, and D1246Y* of multidrug resistance gene 1 (*pfmdr1*) *are shown to be associated with resistance to mefloquine, lumefantrine*, amodiaquine, CQ, and *artemisinin* (Babiker et al., 2001; Humphreys et al., 2007; Somé et al., 2010). Point mutations of the dihydrofolate reductase (*dhfr*) and dihydropteroate synthase (*dhps*) genes, two key enzymes in the folate biosynthesis pathway, mediate resistance to the antifolate drugs sulfadoxine and pyrimethamine (SP), respectively, and have been well described (Gregson and Plowe, 2005). The SNPs have been identified in codons 436, 437, 540, 581, and 613 in the *pfdhps* gene and codons 108, 51, 59, 140, 16, and 164 in the *pfdhfr* gene (Kublin et al., 2002). Mutations in the propeller region of a kelch protein (K13) on *P. falciparum* chromosome 13 (*PF3D7_1343700*) were identified to be associated with artemisinin resistance (Ariey et al., 2014). Understanding the genetic basis of antimalarial drug resistance and identifying specific genetic loci associated with this phenotype are crucial for effective surveillance and containment of resistance.

Microarray-based comparative genomic hybridization, a powerful tool for whole genome analyses and the rapid detection of genomic variation that underlies virulence and disease, provides a robust tool for genome-wide analysis of malaria parasites in diverse settings (Carret et al., 2005). One of the key challenges for microarray analysis is the small amount of genomic DNA obtained from clinical malaria isolates, which is insufficient to be tested in the array. A custom, high-density, NimbleGen microarray covering 33,716 SNPs with high-quality SNPs calls from a wide range of parasite DNA samples, was developed for genome-wide analysis of malaria parasites in different settings (Jacob et al., 2014).

This study addresses the genomic epidemiology of antimalarial drug resistance in *P. falciparum* from the Yunnan and Hainan Provinces in Southern China by using this custom NimbleGen microarray.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

Dried blood spots (DBS) on filter paper (Whatman™ 903, GE Healthcare) were collected from the participants in a therapeutic efficacy study before they received antimalarial drug treatment as well as from the individuals with *P. falciparum* confirmed by microscopy or rapid diagnostic test (RDT) through routine surveillance. Genomic DNA was extracted from the DBS following the manufacturer's instructions (QIAamp 96 DNA Blood Kit, Valencia, CA). Nested polymerase chain reaction (PCR), amplifying the small-subunit rRNA gene of *Plasmodium* spp. (Snounou et al., 1993) was used to confirm the species prior to being tested on the array.

Ethical Considerations

The studies with human subjects were reviewed and approved by the institutional review board of the National Institute of Parasitic Diseases, Chinese Center for Disease Control. In addition, the samples collected from the therapeutic efficacy study were also approved by the WHO Western Pacific Regional Office, and the studies were registered as clinical trials at https://www.anzctr.org.au under the numbers ACTRN12610001008011 and ACTRN12610 001028099. Written informed consent was obtained from patients or guardians.

Quantitative PCR and Whole Gene Amplification (WGA)

Quantitative PCR was used to amplify the *P. falciparum* gene encoding the 18s ribosomal subunit for each sample (Kamau et al., 2011). The total reaction volume was 25 µl, including 2 µl

of sample DNA along with 10 μM probe, 10 μM of each primer, H_2O, and TaqMan universal PCR master mix (containing AmpliTaq Gold DNA Polymerase, dNTPs, and dUTP). The sequences for the primers and probe were Forward - 5′-GTAATTGGAATGATAGGAATTTACAAGGT-3′, Reverse - 5′-TCAACTACGAACGTTTTAACTGCAAC-3′, Probe - 5′-FAM GAACGGGAG GTTAACAA MGB-3′. The PCR conditions were 15 min at 95°C, 15 s at 95°C, and 45 cycles for 1 min at 60°C. The standard curve for DNA quantification was generated and run on each plate as well as a no-DNA control. The standard curve was derived from purified parasite DNA (NF54 strain) and quantified using SYBR Green. This DNA was diluted to 3, 1.5, 0.75, 0.375, 0.188, 0.094, and 0.047 ng/μl, and each standard and sample was tested in duplicate with the final quantity expressed as the mean of both values. The samples with original parasite DNA quantities less than 2 ng were amplified using WGA with the Qiagen REPLI-g mini kit, following the manufacturer's instructions.

SNP Microarray

Parasites were genotyped using a *P. falciparum* SNP microarray. This is a custom NimbleGen 4.2 million probe designed in multiplex format, which comprises 12 identical arrays on each slide (Jacob et al., 2014). One slide is capable of genotyping 33,716 loci within the *P. falciparum* genome. Dual-color labeling was used, and two samples could be hybridized to a single array, yielding 33,716 SNPs for 24 samples in a single experiment. In addition, several slides could be run simultaneously, which made this approach relatively high throughput and low cost. The SNP profiles were examined for the prevalence of validated and candidate molecular markers of drug resistance.

DNA Labeling

Parasite DNA was concentrated using vacuum centrifugation to a volume of 30–50 μl and heat denatured with 1 OD of 65% random nonamers labeled with cy3 or cy5 for 10 min at 98°C. Denatured DNA was chilled on ice for 2 min and then incubated for 2 h at 37°C with 50 units of Klenow fragment and a 50× dNTP mixture. The reaction was terminated with 0.5 M ethylenediamine tetraacetic acid (EDTA), and DNA samples were precipitated with 5 M NaCl and iso-propanol. Labeled DNA was washed 2–3 times with 80% ice-cold ethanol to remove unincorporated dye. After removal of ethanol, the samples were rehydrated in water, and cy3 and cy5 labeled samples were combined for multiplexing. Samples were dried in a SpeedVac on medium heat for 30 min.

Hybridization

Hybridization master mix was prepared with 45.67 μl 2× hybridization buffer, 39.6 μl Denhardt's solution, 18.27 μl hybridization comp and 1.88 μl alignment oligo. It must be mixed very well before use, and 8 μl of hybridization master mix was added to each labeled DNA sample, vortexed well, and quickly spun. The mixture was heat denatured at 95°C for 5 min and stabilized at 42°C prior to loading onto the array. Loaded samples were hybridized on the NimbleGen hybridization station for 16–24 h at 42°C.

Slide Washing and Scanning

Slides were disassembled in a dish containing wash buffer 1 at 42°C and then washed in wash buffer 1, wash buffer 2, and wash buffer 3 for 2 min, 1 min, and 15 s, respectively. The slides were washed and subsequently dried in the Slide Washer 12 Array Processing System. Microarrays were scanned with a NimbleGen MS 200 Microarray Scanner at 2 μm using "auto gain" to automatically adjust the scanning parameters on an individual array basis.

Data Analysis

Spearman's rank correlation was used to evaluate the relationship of DNA quantity between the original extraction and post-WGA and the relationship between parasitemia and the original DNA quantity by SAS software (SAS Institute Inc., Version 9.2, Cary, NC, USA). A P value of <0.05 was used to evaluate the differences with statistical significance. The SNP call rate and SNP call accuracy were calculated based on the intensity of each probe using the heuristic algorithm written in PERL (Jacob et al., 2014) and standard outputs from the Roche NimbleScan (v2.6) software. The samples with the missing calls of the SNPs associated with antimalarial drug resistance are removed in the data analysis. The structure of each parasite population was evaluated by principal component analysis (PCA) using the EIGENSOFT program (Price et al., 2006), and ADMIXTURE (Alexander et al., 2009) was used to calculate maximum likelihood estimates of the most probable number of ancestral populations (K) based on data of the SNPs for the substructure analysis. The SNP data of Southeast Asia samples used in PCA analysis are from the published study (Takala-Harrison et al., 2015).

RESULTS

Sample Information

The DBS were collected from *P. falciparum*–infected patients before antimalarial drug treatment. These cases were diagnosed by microscopy or RDT before sample collection. A total of 256 DBS samples were collected from Yingjian, Menglian, and Tengchong in Yunnan Province from 2009 to 2012 (**Figure 1**). Among them, 65 isolates also had parasitemia data from a level 1 microscopist who was certificated by the WHO. Another 11 DBS from Hainan Province collected in 2007 were only used in subpopulation structure analysis (**Table 1**).

DNA Quantity

After the genomic DNA was extracted, 86.3% (221/256) were successfully quantified of *P. falciparum* genomic DNA, and 35 samples that were tested failed in the quantitative PCR. The threshold for total DNA to be tested in the microarray was 2 ng. The highest DNA quantity in the DBS was 77.51 ng. A total of 64 samples contained more than 2 ng of genomic DNA, which accounted for 25.0%; 25 samples contained DNA between 1 and 2 ng, 75 samples between 0.1 and 1 ng and 57 less than 0.1 ng. The proportion in each range of DNA quantity is shown in **Figure 2**.

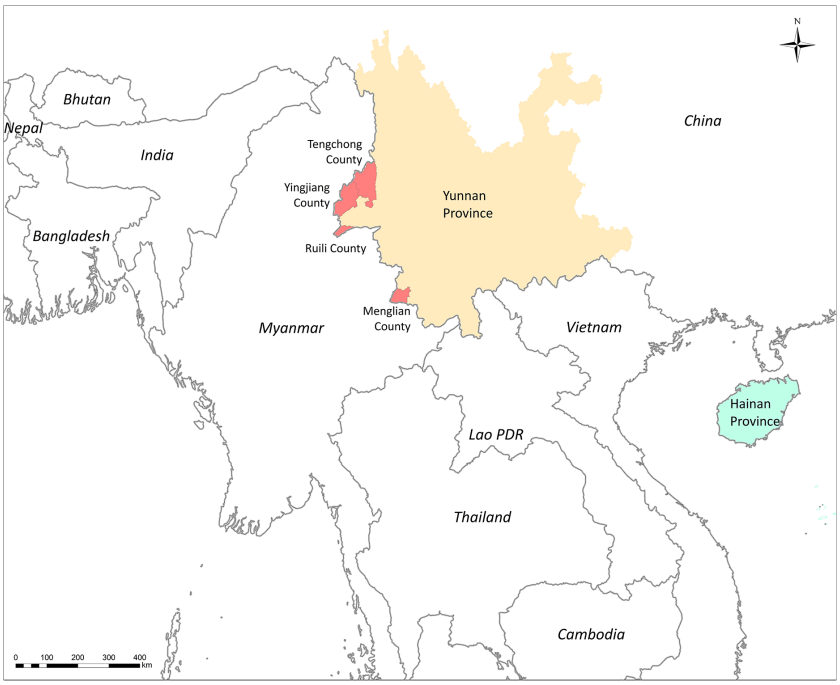


FIGURE 1 | Sites for sample collection in Southern China and countries of samples collection in PCA analysis.

TABLE 1 | Sites and years of collection of dry blood samples of *P. falciparum* in Southern China.

Study site	Year					Total
	2007	2009	2010	2011	2012	
Yingjiang		84	28		54	166
Ruili	16					16
Menglian		44		11		55
Tengchong					19	19
Hainan	11*					
Total		**144**	**28**	**11**	**73**	**256**

The 11 samples from Hainan were only used in the population structure analysis.

At the first step, a total of 48 samples with the original DNA at 0.5–2 ng underwent whole-genome amplification, and 85.6% (43/48) had an increased amount of genomic DNA. The post-WGA DNA amount in 39 samples was more than 2 ng, and the highest post-WGA DNA was 186 times the original DNA amount. Nevertheless, the DNA concentration of post-WGA was not correlated with the original DNA concentration (**Figure 3A**). When the relationship between the parasitemia and DNA quantity was analyzed, there was a significant relation between parasitemia and DNA quantity (r=0.6701, P<0.001) (**Figure 3B**). Second, another 49 samples with the original DNA < 0.5 ng underwent WGA, and 32 samples were increasing the DNA amount to more than 2 ng. Considering the lower of the original DNA of these 49 samples, the relationship analysis between the original DNA and post-WGA DNA did not include these 49 samples.

SNP Call Rate

A total of 135 samples (including 64 with original DNA >2 ng and 71 with post-WGA DNA>2ng) had an average call rate of 44.4% (**Figure 4**). The highest call rate was 89.1%, and the lowest was 18.5%. The SNP call rates were not in a normal distribution by using the Shapiro-Wilk normality test of SAS software (mean: 44.4%, standard deviation: 0.17168, P<0.01). The samples with the highest call rate also had the highest parasite DNA level. In addition, there was a strong correlation between DNA quantity and call rate (data not shown), which demonstrated that DNA quantity was a good predictor of the number of SNPs called.

Molecular Markers

A total of 27 SNPs related to antimalarial drug resistance were tested through the SNP microarray. These SNPs were from *pfcrt*, *pfmdr1*, *pfdhfr*, *phdhps*, and another four SNPs (MAL10:688956, MAL13:1718319, MAL13:1719976, MAL14:718269) located on chromosomes 10, 13, and 14, which were associated with delayed parasite clearance time. The location of each SNP is shown in the **Supplementary File 1**.

In total, eight codons of *pfcrt* were tested, and no mutations were identified at codons 74, 75, and 371 (**Figure 5**). All 66 samples tested successfully were 100.0% mutant at *pfcrt* 356 (66/66), followed by codons 326 (91.2%, 31/34), 271 (57.1%, 24/42), 76 (39.3%, 11/27), and 220 (9.7%, 3/31). All the SNP mutations involving codons N86Y, Y184F, S1034C, N1042D, and D1246Y of *pfmdr1* were wild type.

A total of 10 SNPs of *pfdhfr* and *pfdhps* associated with antifolate drugs, located on chromosomes 4 and 8, respectively,

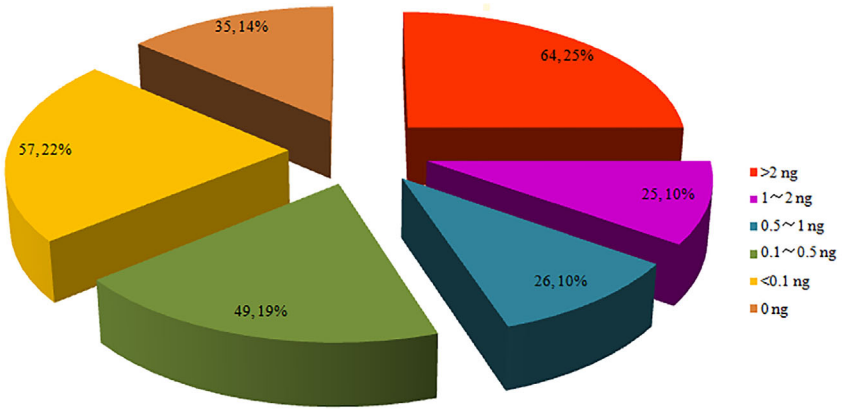


FIGURE 2 | Proportion of parasite genomic DNA evaluated by 18s rRNA qPCR.

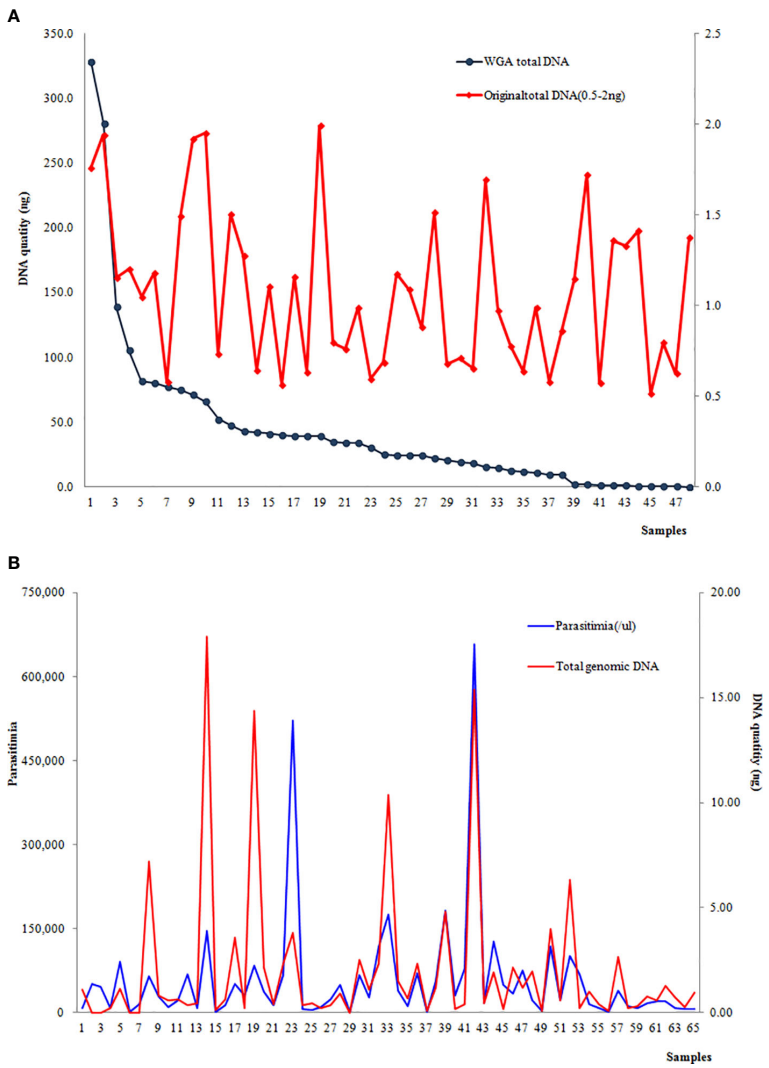


FIGURE 3 | **(A)** Correlation between the DNA concentration of the original extraction and post-WGA; **(B)** Correlation between parasitemia and the original DNA quantity.

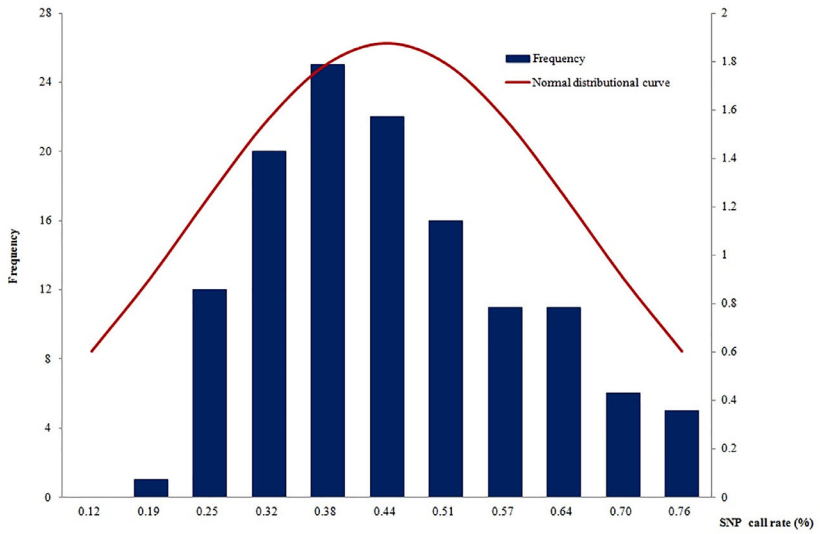


FIGURE 4 | The SNP call rates were not in a normal distribution (mean: 44.38%, standard deviation: 0.17168, $P<0.01$).

were tested on the microarray. The highest prevalence of *pfdhfr* was at codon 108 (96.7%, 29/30), followed by codons 59 (40.0%, 16/40), 51 (16.0%, 4/25), and 164 (6.3%, 3/48). No mutations were identified at codon 51. The prevalence of *pfdhps* mutants was consistent with that of *pfdhfr*. The prevalence of *pfdhps* mutation at codons 436, 437, 540, and 581 was 98.5%, 56.3%, 55.6%, and 71.2%, respectively (**Figure 5**).

Candidate SNPs on chromosomes 10, 13, and 14 that were associated with delayed parasite clearance showed low prevalence of mutants. The SNP at MAL13:1719976 showed a 100.0% mutation rate. Nevertheless, the other three had low mutant prevalence or were wild type.

Population Structure

All the samples with successful SNP calls were used to evaluate the structure of each parasite population. After the SNP data were filtered with the threshold of missing data <50%, 686 isolates, including 135 from this study and 551 from the published data (Takala-Harrison et al., 2015), along with 570 SNP data by using the same NimbleGen microarray were used for PCA. PCA illustrates the first PC on the horizontal axis and the second and third PC on the vertical axis, respectively (**Figure 6**). The samples are colored by geographical location stratified by country in Southeast Asia. PCA results indicate a clear distinction between the isolates. Parasites from Southern China are clustered and separated from the isolates from western Cambodia, northwest Thailand, southern Myanmar, Vietnam, and Bangladesh in Southeast Asia (Takala-Harrison et al., 2015) (**Figure 6**).

The substructure of 135 samples of different origins in Southern China, including Yunnan and Hainan Provinces, was evaluated by ADMIXTURE. The parasite populations were designed on the basis of their geographic origins with K=5. Each vertical line represents a single sample with color denoting the origin proportion in that sample (**Figure 7**). Admixture estimates were computed using the parameters over a varying

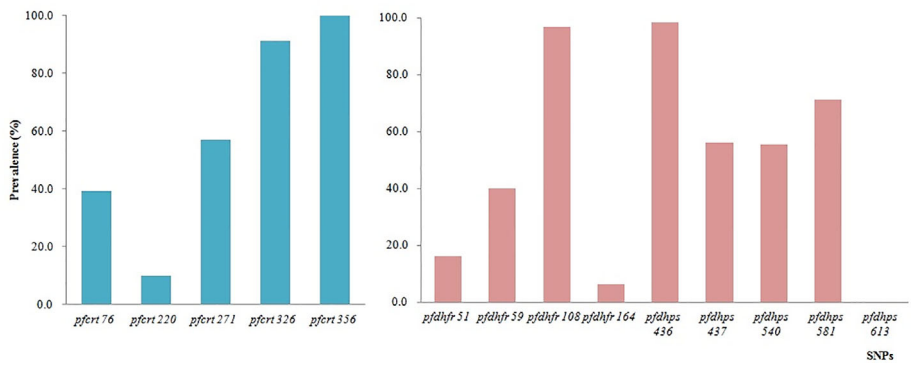


FIGURE 5 | Prevalence of constructed haplotypes of SNPs in *pfcrt*, *pfdhfr*, and *pfdhps* at different codons.

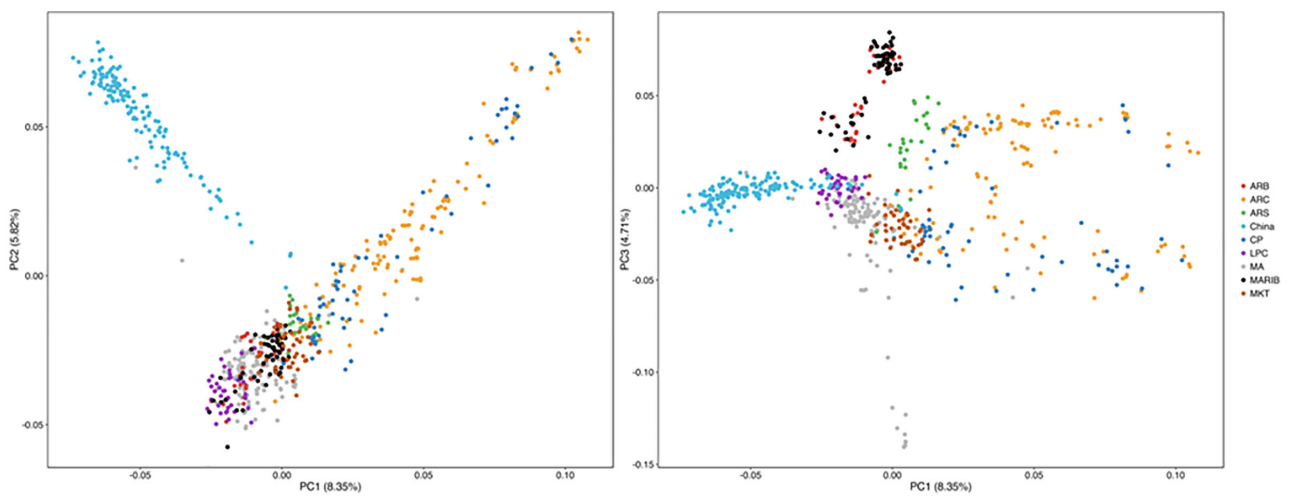


FIGURE 6 | Parasites from Southern China were clustered with and separated from parasites from Southeast Asia by PCA. ARB/MARIB: Bangladesh; ARC/CP: Western Cambodia; ARS: Northwest Thailand; LPC = Laos; MA: Vietnam; MKT: Southern Myanmar.

number of clusters (K) ranging from 2 to 10 with 10 technical replicates each. The optimal K value was determined by maximizing the log-likelihood across replicates for a single K value and minimizing the cross-validation error between different K values (**Supplementary File 2**). In this study, the optimal K was determined as 2, which indicates that a total of 135 samples from Southern China were mainly divided into two subgroups. The parasites from the different sites in Yunnan Province bordering with Myanmar belong to one group, and the parasites from Hainan Province are slightly separated from parasites from Yunnan Province (**Figure 7**).

DISCUSSION

Malaria was one of the most serious infectious diseases in the last century and was endemic in central and Southern China (Tang et al., 1997). However, China has made great contributions toward global malaria control in the past 40 years, and in 2010, China launched the National Malaria Elimination Action Plan 2010–2020 with an ultimate goal to interrupt local malaria transmission by 2020 (Feng et al., 2016; Yang and Zhou, 2016; Feng et al., 2020). ACTs, considered to be the best therapy for falciparum malaria in the world, have contributed to significant decreases in case numbers and deaths and are crucial to the success of control and elimination programs (Nosten and White, 2007; Dondorp et al., 2009). Recently, the emergence and spread of artemisinin resistance in *P. falciparum* poses a threat to malaria control and eradication goals in the GMS, where the resistance has emerged independently and has spread (Henriques et al., 2013; Ashley et al., 2014; Huang et al., 2015; Takala-Harrison et al., 2015; Ménard et al., 2016). Parasites in Southern China are reported to show decreased *in vitro* sensitivity to artemisinin and delayed parasite clearance time of artemisinin (Huang et al., 2015; Wang et al., 2015).

In this study, parasites collected from Southern China were genotyped using an SNP microarray, including 33,716 loci within the *P. falciparum* genome. This new SNP microarray was developed

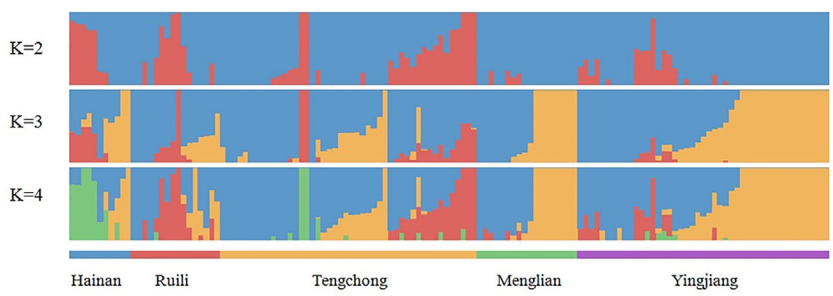


FIGURE 7 | Population structure of the sample set, analyzed on the basis of geography, using ADMIXTURE. Parasites from the different sites in Yunnan Province along the border areas and Hainan Province were slightly separated.

by using DBS, which did not have a sufficiently high concentration of DNA for sequencing (Jacob et al., 2014). We chose 64 samples with original DNA content of more than 2 ng and 71 samples of post-WGA DNA >2 ng. The average SNP call rate was 44.4% using the DBS in this study, and this was not as high as for venous blood samples (Jacob et al., 2014). WGA is one way to increase the total quantity of whole DNA, including human and parasite DNA (Nakayama et al., 2008; Chueh et al., 2011). Interestingly, the DNA concentration post-WGA was not correlated with the original DNA concentration in this study. The ratio of human DNA to parasite DNA prior to amplification may contribute to this discrepancy as a majority of the DNA sequences were from humans. Additionally, we observed a significant relation between the level of parasitemia and DNA quantity.

A total of 27 validated and candidate SNPs associated with antimalarial drug resistance were tested. Parasites in Southern China show a moderate prevalence of *pfcrt* mutation with a prevalence of 39.3%, which is much lower than that in another study that reported 100.0% mutant haplotype of *pfcrt* CIETS (Bai et al., 2018). However, the prevalence of *pfcrt* at Q271E, N326S, and I356L remain high in this study. Interestingly, the mutations of *pfmdr1* involving codons N86Y, Y184F, S1034C, N1042D, and D1246Y were all of wild type. CQ was used to treat falciparum malaria in China in the last century until it was withdrawn and replaced in the 1970s as a result of the CQ resistance emerging in Southern China according to *in vivo* and *in vitro* testing (Huang et al., 1988; Hong-Ping et al., 1998; Liu et al., 2005). This decreased prevalence of the *pfcrt* K76T marker may be associated with the cessation of CQ use against *P. falciparum* malaria in Southern China (Wang et al., 2005; Laufer et al., 2010).

The prevalence of *pfdhfr* and *pfdhps* remains at a high level even though antifolate drugs have not been used in China for many years, which is consistent with other studies (Huang et al., 2012a; Huang et al., 2012b; Bai et al., 2018). Pyrimethamine was used for the radical treatment of *P. vivax* in combination with primaquine in China around 40 years ago (unpublished data). Additionally, pyrimethamine plus primaquine has always been recommended as prophylaxis for specific populations, and pyrimethamine was added to salt for prophylaxis in China in the 1980s (Wang, 1984). The mechanism of action of *pfdhfr* and *pfdhps* mutations on the resistance to SP drugs has not been well described.

When the SNP microarray was developed, the K13 gene associated with artemisinin resistance was not identified. Therefore, only four candidate SNPs associated with delayed parasite clearance time were included in the microarray. Only one locus, MAL13:1719976, showed 100.0% mutation in all the samples, and the other three had low levels of mutation or were wild type. These results have provided further evidence of the decreased sensitivity to artemisinin and delayed parasite clearance time of artemisinin in Southern China, which was identified by our other published study (Huang et al., 2015).

According to the analysis of population structure, the parasites from Southern China represent a clear distinct cluster and little connection with the parasites from other countries in the GMS, which is consistent with the findings of other studies (Wei et al., 2015; Zhu et al., 2016; Ye et al., 2019). Although the population diversity of *P. falciparum* is high in the GMS, the parasites from Southern China are relatively low diversity. Human migration may be another cause for parasite population evolution and genomic diversity. Yunnan Province has a long border with Myanmar, but the human population movement is relatively low because of geographical barriers and some regions having political issues (Wang et al., 2016; Xu and Liu, 2016; Zhou et al., 2016; Chen et al., 2018). Hainan Province, geographically separated from the Chinese mainland, provides a natural barrier to parasite migration and spread. The population structure analysis is a useful tool that allows targeting of populations with low migration and low diversity in the malaria elimination stage and tracking of the origin of imported parasites in the postelimination stage.

LIMITATIONS

In this study, we used samples of DBS from the field that had low quantity genomic DNA; even though we used WGA to increase the DNA amount, the total DNA were still lower than venous blood samples. In addition, the SNP call rates of DBS samples were not as good as those of venous blood samples. The DNA microarray could only test known SNPs. We did not test the SNPs of K13 using this microarray because K13 was not identified when the chip was developed. Although this custom array is not available now, some new microarrays have been developed, for example, Illumina Bead Chip microarray, protein array, or peptide array, which provide a high throughout and a powerful platform in malaria field (Bailey et al., 2020; Zhao et al., 2020).

CONCLUSIONS

Parasites from Southern China were clustered and separated from those from Southeast Asia although parasites from Yunnan Province were a substructure of those from Hainan Province. This study provides a population-level genetic framework for investigating the biological origins of antimalarial drug resistance.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. Data from the JID paper that was used for comparison with the China samples is publicly available at: NIH Gene Expression Omnibus (www.ncbi.nlm.gov/geo/) (Accession number: GSE100704).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the National Institute of Parasitic Diseases, Chinese

Center for Disease Control. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

FH, CJ, and CP conceived and designed the study. CJ, ST-H, and FH carried out the data analysis. CJ, MA, and FH conducted the laboratory work. H-LY and HL collected the field samples. Z-GX, S-SZ, and L-HT contributed to the manuscript edits. FH drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The study was supported by the Natural Science Foundation of Shanghai (No. 18ZR1443400), WHO Mekong Malaria Programme (WP/08/MVP/000512 and WP/10/MVP/005837), and the Howard Hughes Medical Institute and fellowship supported by Shanghai government. R01 (grant number R01AI101713) from the National Institute of Allergy and Infectious Diseases at the National Institutes of Health paid for the genotyping of the samples from the JID paper.

ACKNOWLEDGMENTS

We thank all the participants who provided blood samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2020.610985/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi: 10.1101/gr.094052.109

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A. C., Khim, N., et al. (2014). A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* 505 (7481), 50–55. doi: 10.1038/nature12876

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., et al. (2014). Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* 371 (5), 411–423. doi: 10.1056/NEJMoa1314981

Babiker, H. A., Pringle, S. J., Abdel-Muhsin, A., Mackinnon, M., Hunt, P., and Walliker, D. (2001). High-level chloroquine resistance in Sudanese isolates of *Plasmodium falciparum* is associated with mutations in the chloroquine resistance transporter gene pfcrt and the multidrug resistance Gene pfmdr1. *J. Infect. Dis.* 183 (10), 1535–1538. doi: 10.1086/320195

Bai, Y., Zhang, J., Geng, J., Xu, S., Deng, S., Zeng, W., et al. (2018). Longitudinal surveillance of drug resistance in *Plasmodium falciparum* isolates from the China-Myanmar border reveals persistent circulation of multidrug resistant parasites. *Int. J. Parasitol. Drugs Drug Resist.* 8 (2), 320–328. doi: 10.1016/j.ijpddr.2018.05.003

Bailey, J. A., Berry, A. A., Travassos, M. A., Ouattara, A., Boudova, S., Dotsey, E. Y., et al. (2020). Microarray analyses reveal strain-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 variants following natural infection and vaccination. *Sci. Rep.* 10 (1), 3952. doi: 10.1038/s41598-020-60551-z

Carret, C. K., Horrocks, P., Konfortov, B., Winzeler, E., Qureshi, M., Newbold, C., et al. (2005). Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. *Mol. Biochem. Parasitol.* 144 (2), 177–186. doi: 10.1016/j.molbiopara.2005.08.010

Chen, T. M., Zhang, S. S., Feng, J., Xia, Z. G., Luo, C. H., Zeng, X. C., et al. (2018). Mobile population dynamics and malaria vulnerability: a modelling study in the China-Myanmar border region of Yunnan Province, China. *Infect. Dis. Poverty* 7 (1), 36. doi: 10.1186/s40249-018-0423-6

Chueh, B. H., Li, C. W., Wu, H., Davison, M., Wei, H., Bhaya, D., et al. (2011). Whole gene amplification and protein separation from a few cells. *Anal. Biochem.* 411 (1), 64–70. doi: 10.1016/j.ab.2010.12.028

Dahlstrom, S., Ferreira, P. E., Veiga, M.II, Sedighi, N., Wiklund, L., Martensson, A., et al. (2009). *Plasmodium falciparum* multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *J. Infect. Dis.* 200 (9), 1456–1464. doi: 10.1086/606009

Djimde, A., Doumbo, O. K., Cortese, J. F., Kayentao, K., Doumbo, S., Diourte, Y., et al. (2001). A molecular marker for chloroquine-resistant falciparum malaria. *N. Engl. J. Med.* 344 (4), 257–263. doi: 10.1056/NEJM200101253440403

Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., et al. (2009). Artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* 361 (5), 455–467. doi: 10.1056/NEJMoa0808859

Feng, J., Liu, J., Feng, X., Zhang, L., Xiao, H., and Xia, Z. (2016). Towards Malaria Elimination: Monitoring and Evaluation of the "1-3-7" Approach at the China-Myanmar Border. *Am. J. Trop. Med. Hyg.* 95 (4), 806–810. doi: 10.4269/ajtmh.15-0888

Feng, X., Levens, J., and Zhou, X. N. (2020). Protecting the gains of malaria elimination in China. *Infect. Dis. Poverty* 9 (1), 43. doi: 10.1186/s40249-020-00661-y

Ferreira, P. E., Holmgren, G., Veiga, M.II, Uhlen, P., Kaneko, A., and Gil, J. P. (2011). PfMDR1: mechanisms of transport modulation by functional polymorphisms. *PloS One* 6 (9), e23875. doi: 10.1371/journal.pone.0023875

Gregson, A., and Plowe, C. V. (2005). Mechanisms of resistance of malaria parasites to antifolates. *Pharmacol. Rev.* 57 (1), 117–145. doi: 10.1124/pr.57.1.4

Henriques, G., Martinelli, A., Rodrigues, L., Modrzynska, K., Fawcett, R., Houston, D. R., et al. (2013). Artemisinin resistance in rodent malaria–mutation in the AP2 adaptor mu-chain suggests involvement of endocytosis and membrane protein trafficking. *Malar. J.* 12, 118. doi: 10.1186/1475-2875-12-118

Holmgren, G., Hamrin, J., Svard, J., Martensson, A., Gil, J. P., and Bjorkman, A. (2007). Selection of pfmdr1 mutations after amodiaquine monotherapy and amodiaquine plus artemisinin combination therapy in East Africa. *Infect. Genet. Evol.* 7 (5), 562–569. doi: 10.1016/j.meegid.2007.03.005

Hong-Ping, G., Yao, Z., Qian, D., Jin-Hua, Y., Kai-Xun, J., and Min, Z. (1998). The in vivo sensitivity of *Plasmodium falciparum* to chloroquine in the Red River basin, Yunnan, China. *Southeast Asian J. Trop. Med. Public Health* 29 (4), 692–695.

Huang, O. L., Ouyang, W. C., Zhou, J. X., Wu, Z., Zhang, K. Y., Huang, J. K., et al. (1988). Effectiveness of amodiaquine, sulfadoxine-pyrimethamine, and combinations of these drugs for treating chloroquine-resistant falciparum malaria in Hainan Island, China. *Bull. World Health Organ* 66 (3), 353–358.

Huang, F., Tang, L., Yang, H., Zhou, S., Liu, H., Li, J., et al. (2012a). Molecular epidemiology of drug resistance markers of *Plasmodium falciparum* in Yunnan Province, China. *Malar. J.* 11, 243. doi: 10.1186/1475-2875-11-243

Huang, F., Tang, L., Yang, H., Zhou, S., Sun, X., and Liu, H. (2012b). Therapeutic efficacy of artesunate in the treatment of uncomplicated *Plasmodium falciparum* malaria and anti-malarial, drug-resistance marker polymorphisms in populations near the China-Myanmar border. *Malar. J.* 11, 278. doi: 10.1186/1475-2875-11-278

Huang, F., Takala-Harrison, S., Jacob, C. G., Liu, H., Sun, X., Yang, H., et al. (2015). A single mutation in K13 predominates in Southern China and is associated with delayed clearance of *Plasmodium falciparum* following

artemisinin treatment. *J. Infect. Dis.* 212 (10), 1629–1635. doi: 10.1093/infdis/jiv249

Humphreys, G. S., Merinopoulos, I., Ahmed, J., Whitty, C. J., Mutabingwa, T. K., Sutherland, C. J., et al. (2007). Amodiaquine and artemether-lumefantrine select distinct alleles of the *Plasmodium falciparum* mdr1 gene in Tanzanian children treated for uncomplicated malaria. *Antimicrob. Agents Chemother.* 51 (3), 991–997. doi: 10.1128/AAC.00875-06

Jacob, C. G., Tan, J. C., Miller, B. A., Tan, A., Takala-Harrison, S., Ferdig, M. T., et al. (2014). A microarray platform and novel SNP calling algorithm to evaluate *Plasmodium falciparum* field samples of low DNA quantity. *BMC Genomics* 15, 719. doi: 10.1186/1471-2164-15-719

Kamau, E., Tolbert, L. S., Kortepeter, L., Pratt, M., Nyakoe, N., Muringo, L., et al. (2011). Development of a highly sensitive genus-specific quantitative reverse transcriptase real-time PCR assay for detection and quantitation of plasmodium by amplifying RNA and DNA of the 18S rRNA genes. *J. Clin. Microbiol.* 49 (8), 2946–2953. doi: 10.1128/JCM.00276-11

Kublin, J. G., Dzinjalamala, F. K., Kamwendo, D. D., Malkin, E. M., Cortese, J. F., Martino, L. M., et al. (2002). Molecular markers for failure of sulfadoxine-pyrimethamine and chlorproguanil-dapsone treatment of *Plasmodium falciparum* malaria. *J. Infect. Dis.* 185 (3), 380–388. doi: 10.1086/338566

Lakshmanan, V., Bray, P. G., Verdier-Pinard, D., Johnson, D. J., Horrocks, P., Muhle, R. A., et al. (2014). A critical role for PfCRT K76T in *Plasmodium falciparum* verapamil-reversible chloroquine resistance. *EMBO J.* 24 (13), 2294–2305. doi: 10.1038/sj.emboj.7600681

Laufer, M. K., Takala-Harrison, S., Dzinjalamala, F. K., Stine, O. C., Taylor, T. E., and Plowe, C. V. (2010). Return of chloroquine-susceptible falciparum malaria in Malawi was a reexpansion of diverse susceptible parasites. *J. Infect. Dis.* 202 (5), 801–808. doi: 10.1086/655659

Liu, D. Q., Feng, X. P., Yang, H. L., Lin, S. G., Chen, W. J., and Yang, P. F. (2005). Fluctuation in the resistance of *Plasmodium falciparum* to chloroquine in China. *Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi* 23 (1), 27–31.

Ménard, D., Khim, N., Beghain, J., Adegnika, A. A., Shafiul-Alam, M., Amodu, O., et al. (2016). A Worldwide Map of *Plasmodium falciparum* K13-Propeller Polymorphisms. *N. Engl. J. Med.* 374 (25), 2453–2464. doi: 10.1056/NEJMoa1513137

Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., et al. (2008). The Whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. *DNA Res.* 15 (4), 185–199. doi: 10.1093/dnares/dsn011

NHC (2016). *Technical regulations for application of antimalarials* (Beijing: National Health Commission of the People's Republic of China).

Nosten, F., and White, N. J. (2007). Artemisinin-based combination treatment of falciparum malaria. *Am. J. Trop. Med. Hyg.* 77 (6 Suppl), 181–192. doi: 10.4269/ajtmh.2007.77.181

Picot, S., Olliaro, P., de Monbrison, F., Bienvenu, A. L., Price, R. N., and Ringwald, P. (2009). A systematic review and meta-analysis of evidence for correlation between molecular markers of parasite resistance and treatment outcome in falciparum malaria. *Malar. J.* 8, 89. doi: 10.1186/1475-2875-8-89

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–909. doi: 10.1038/ng1847

Snounou, G., Viriyakosol, S., Zhu, X. P., Jarra, W., Pinheiro, L., do Rosario, V. E., et al. (1993). High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. *Mol. Biochem. Parasitol.* 61 (2), 315–320. doi: 10.1016/0166-6851(93)90077-b

Somé, A. F., Séré, Y. Y., Dokomajilar, C., Zongo, I., Rouamba, N., Greenhouse, B., et al. (2010). Selection of known *Plasmodium falciparum* resistance-mediating polymorphisms by artemether-lumefantrine and amodiaquine-sulfadoxine-pyrimethamine but not dihydroartemisinin-piperaquine in Burkina Faso. *Antimicrob. Agents Chemother.* 54 (5), 1949–1954. doi: 10.1128/AAC.01413-09

Srimuang, K., Miotto, O., Lim, P., Fairhurst, R. M., Kwiatkowski, D. P., Woodrow, C. J., et al. (2016). Analysis of anti-malarial resistance markers in pfmdr1 and pfcrt across Southeast Asia in the Tracking Resistance to Artemisinin Collaboration. *Malar. J.* 15 (1), 541. doi: 10.1186/s12936-016-1598-6

Takala-Harrison, S., Jacob, C. G., Arze, C., Cummings, M. P., Silva, J. C., Dondorp, A. M., et al. (2015). Independent emergence of artemisinin resistance mutations among *Plasmodium falciparum* in Southeast Asia. *J. Infect. Dis.* 211 (5), 670–679. doi: 10.1093/infdis/jiu491

Tang, L. H., Qian, H. L., Cui, G., Shang, L. Y., Tang, L. Y., Jiang, M. G., et al. (1997). Study of simplified measures for malaria surveillance in the late consolidation phase in China. *Southeast Asian J. Trop. Med. Public Health* 28 (1), 4–115.

Tu, Y. (2011). The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.* 17 (10), 1217–1220. doi: 10.1038/nm.2471

Vinayak, S., Alam, M. T., Sem, R., Shah, N. K., Susanti, A.II, Lim, P., et al. (2010). Multiple genetic backgrounds of the amplified *Plasmodium falciparum* multidrug resistance (pfmdr1) gene and selective sweep of 184F mutation in Cambodia. *J. Infect. Dis.* 201 (10), 1551–1560. doi: 10.1086/651949

Wang, X., Mu, J., Li, G., Chen, P., Guo, X., Fu, L., et al. (2005). Decreased prevalence of the *Plasmodium falciparum* chloroquine resistance transporter 76T marker associated with cessation of chloroquine use against *P. falciparum* malaria in Hainan, People's Republic of China. *Am. J. Trop. Med. Hyg.* 72 (4), 410–414.

Wang, Z., Wang, Y., Cabrera, M., Zhang, Y., Gupta, B., Wu, Y., et al. (2015). Artemisinin resistance at the China-Myanmar border and association with mutations in the K13 propeller gene. *Antimicrob. Agents Chemother.* 59 (11), 6952–6595. doi: 10.1128/AAC.01255-15

Wang, R. B., Dong, J. Q., Xia, Z. G., Cai, T., Zhang, Q. F., Zhang, Y., et al. (2016). Lessons on malaria control in the ethnic minority regions in Northern Myanmar along the China borde-2014. *Infect. Dis. Poverty* 5 (1), 955. doi: 10.1186/s40249-016-0191-0

Wang, R. Z. (1984). Efficiency of pyrimethamine salt for *P. vivax* prophylaxis. *Railway Med.* 4, 246–247.

Wei, G., Zhang, L., Yan, H., Zhao, Y., Hu, J., and Pan, W. (2015). Evaluation of the population structure and genetic diversity of *Plasmodium falciparum* in southern China. *Malar. J.* 14, 283. doi: 10.1186/s12936-015-0786-0

WHO (2015). *Strategy for malaria elimination in the GMS (2015–2030)* (Geneva: World Health Organization).

WHO (2019). *World malaria report 2019* (Geneva: World Health Organization).

Xu, J. W., and Liu, H. (2016). The relationship of malaria between Chinese side and Myanmar's five special regions along China-Myanmar border: a linear regression analysis. *Malar. J.* 15, 368. doi: 10.1186/s12936-016-1413-4

Yang, W. Z., and Zhou, X. N. (2016). New challenges of malaria elimination in China. *Zhonghua Yu Fang Yi Xue Za Zhi* 50 (4), 289–291. doi: 10.3760/cma.j.issn.0253-9624.2016.04.001

Ye, R., Tian, Y., Huang, Y., Zhang, Y., Wang, J., Sun, X., et al. (2019). Genome-wide analysis of genetic diversity in *Plasmodium falciparum* isolates from China-Myanmar border. *Front. Genet.* 10, 1065. doi: 10.3389/fgene.2019.01065

Zhao, Y., Hosking, C., Cunningham, D., Langhorne, J., and Lin, J. W. (2020). Transcriptome analysis of blood and spleen in virulent and avirulent mouse malaria infection. *Sci. Data* 7 (1), 253. doi: 10.1038/s41597-020-00592-1

Zhou, G., Lo, E., Zhong, D., Wang, X., Wang, Y., Malla, S., et al. (2016). Impact of interventions on malaria in internally displaced persons along the China-Myanmar border: 2011-2014. *Malar. J.* 15, 471. doi: 10.1186/s12936-016-1512-2

Zhu, X., Zhao, Z., Feng, Y., Li, P., Liu, F., Liu, J., et al. (2016). Genetic diversity of the *Plasmodium falciparum* apical membrane antigen I gene in parasite population from the China-Myanmar border area. *Infect. Genet. Evol.* 39, 155–162. doi: 10.1016/j.meegid.2016.01.021

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.

Copyright © 2021 Huang, Jacob, Takala-Harrison, Adams, Yang, Liu, Xia, Zhou, Tang and Plowe. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cellular and Infection Microbiology

METHODS
published: 11 January 2021
doi: 10.3389/fcimb.2020.620419

Sensitive Immunoassay Detection of *Plasmodium* Lactate Dehydrogenase by Inductively Coupled Plasma Mass Spectrometry

Jianbing Mu[1], *Lee L. Yu*[2] *and Thomas E. Wellems*[1]*

[1] Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, United States, [2] Chemical Sciences Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, United States

Rapid, reliable, and sensitive detection of *Plasmodium* infection is central to malaria control and elimination. Many Malaria Rapid Diagnostic Tests (RDTs) developed for this purpose depend upon immunoassays that can be improved by advances in bound antibody sensor technology. In a previous study, immuno-polymerase chain reaction (PCR) was shown to provide highly sensitive detection of *Plasmodium falciparum* lactate dehydrogenase (PfLDH) in monoclonal antibody (mAb) sandwich assays. Here, we show comparably high immunoassay sensitivity by inductively coupled plasma mass spectrometry (ICP-MS) detection of gold nanoparticles (AuNPs). Following capture of PfLDH with the primary mAb and binding of the AuNP-labeled detection mAb, ICP-MS signals from the AuNPs provided quantitative measures of recombinant PfLDH test dilutions and *P. falciparum*-infected erythrocytes. A detection limit of 1.5 pg/mL was achieved with the PfLDH protein. Parasitemia in cultures of *P. falciparum*-infected erythrocytes could be detected to a lower limit of 1.6 parasite/μl (p/μl) for early ring-stage forms and 0.3 p/μl for mixed stages including mature trophozoites and schizont-stages. These results show that ICP-MS detection of AuNPs can support highly sensitive and accurate detection of *Plasmodium* infection.

Keywords: Gold nanoparticles, malaria, diagnostic tests, antigen-based detection, ICP-MS

OPEN ACCESS

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

Reviewed by:
Qingfeng Zhang,
Tongji University, China
Kenji Hirayama,
Nagasaki University, Japan
Gaoqian Feng,
Burnet Institute, Australia

****Correspondence:***
Thomas E. Wellems
TWELLEMS@niaid.nih.gov

Specialty section:
This article was submitted to Parasite and Host, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *22 October 2020*
Accepted: *24 November 2020*
Published: *11 January 2021*

Citation:
Mu J, Yu LL and Wellems TE (2021) Sensitive Immunoassay Detection of Plasmodium Lactate Dehydrogenase by Inductively Coupled Plasma Mass Spectrometry. Front. Cell. Infect. Microbiol. 10:620419. doi: 10.3389/fcimb.2020.620419

INTRODUCTION

Asymptomatic and submicroscopic *Plasmodium* infections that go undetected in endemic populations and sustain transmission through mosquitoes are a continuing impediment to malaria elimination (Wu et al., 2015). Rapid and reliable diagnostic tests are needed that can detect these infections and be widely deployed for use at the primary healthcare level (Gitta and Kilian, 2020). Microscopy of Giemsa-stained blood smears falls short of these needs, not only because of the expertise and time it demands, but also by its detection threshold, estimated to be 50 to 500 parasites/μl (p/μl) under field conditions (Poschl et al., 2010). Highly sensitive molecular tests based on polymerase chain reaction (PCR) (Poschl et al., 2010), real-time quantitative PCR (qPCR) (Grignard et al., 2020), reverse transcription polymerase chain reaction (RT-PCR)

(Bourgeois et al., 2010), loop-mediated isothermal amplification (LAMP) (Han et al., 2007; Kongkasuriyachai et al., 2017; Mohon et al., 2019) and a clustered, regularly-interspaced palindromic repeat (CRISPR)-based assay (Cunningham et al., 2020) are much more sensitive, detecting parasitemia as low as 2 to 5 p/μl. However, the requirements of nucleic acid extraction, amplification, and detection often limit the use of these methods in resource-poor endemic settings.

Between the relatively low sensitivity of microscopy and the high sensitivity of nucleotide sequence detection methods, a mid-range of sensitivity is offered by Rapid Diagnostic Tests (RDTs), of which some can detect as few as 10 p/μl (Hemben et al., 2018). Overall, an estimated 1.92 billion RDTs were distributed between 2010 and 2017 (Slater et al., 2015; Wu et al., 2015). Antigens detected by these RDTs include the *P. falciparum* histidine-rich protein II and III (PfHRP-II, PfHRP-III), lactate dehydrogenase of *P. falciparum* (PfLDH) as well as other *Plasmodium* species, or a *Plasmodium* aldolase, all of which are abundantly produced by the parasites during the erythrocytic cycle. Although the tests can provide rapid results and require no specific training or equipment, the performances of RDTs call for improvements on their mid-range sensitivity (Das et al., 2018) and false-negative results, which can result from loss of PfHRP-II/III expression from some *P. falciparum* strains (Gatton et al., 2017; Verma et al., 2018; Bosco et al., 2020) or bloodstream parasitemias below the detection threshold (Bell et al., 2006).

To improve the performance of RDTs, research efforts in recent years have included searches for alternative parasite antigens (Mu et al., 2017), new detection methods for hemozoin from parasite digestion of hemoglobin (Peng et al., 2014; Rifaie-Graham et al., 2019), and improved biosensors and immunosensors (Krampa et al., 2020). Electrochemical techniques, with labeled amperometric as well as label-free impedimetric strategies, have been applied to malaria diagnostic studies and outperformed optical methods (Sharma et al., 2011; Low et al., 2019; Obisesan et al., 2019). A sensitive, magnetic nanoparticle (MNP) labeled immunosensor has been shown to detect PfHRP-II with a limit of detection (LOD) of 360 pg/mL (Castilho Mde et al., 2011). Aptamer-based sensors targeting PfLDH demonstrated high sensitivity and specificity with LOD measuring 120.1 fmol/L (4 pg/ml) (Lee et al., 2012). Dirkzwager et al. (Dirkzwager et al., 2016) used aptamer functionalized microbeads to measure the enzymatic activity of recombinant PfLDH and showed a LOD of 4.9 ng/ml. Further, a portable microfluidic biosensor, integrated with the aptamer-tethered enzyme capture (APTEC) assay, was developed for highly specific detection of *P. falciparum* in cultures and clinical samples (Fraser et al., 2018). Receptor-target binding can be detected in label-free formats by piezoelectric immunosensors (Sharma et al., 2011) or indicator displacement assays (Chakma et al., 2016). Although these methods are considered to have advantages for field applications, their sensitivity (12.0 ng/ml) needs to be improved further.

Gold nanoparticles (AuNPs) offer highly useful physicochemical properties including high density, good biocompatibility, stability, and catalysis activities, which have spurred their application in bioassays and diagnostics, particularly to meet the needs of points of care (Cordeiro et al., 2016). Indeed, for malaria parasite detection, assays with AuNPs have been developed for rapid, simple, and cost-effective detection of *P. falciparum* infections, although the estimated LOD for these assays was poor (2.4 μg/ml for PfHsp70 and PfHSP90; ≈1,000 p/μl for PfLDH) (Castilho Mde et al., 2011; Guirgis et al., 2012; Jeon et al., 2013). To improve on the detection sensitivity of biosensors for AuNPs, non-optical bioassays have been developed with piezoelectric and electrochemical biosensors (Jiang et al., 2018). AuNP-labeled sandwich assay on the surface of a screen-printed gold electrode provided improved sensitivity of PfHRP2 detection to 36 pg/ml (Hemben et al., 2017) and of PfLDH to 19.0 pg/ml (Hemben et al., 2018). AuNPs have served for an electrochemical sensor in the detection of malaria parasites in clinical samples (Obisesan et al., 2019).

Inductively coupled plasma mass spectrometry (ICP-MS), is a highly sensitive technique for element and isotope analysis. The high-temperature ionization characteristics of the inductively coupled plasma (ICP), together with the sensitive and fast scanning modern mass analyzer, have boosted the capability of ICP-MS detection for biomolecules with metallic nanoparticle (NP) labels. In biomedical diagnostics, ICP-MS measurement of AuNP labels for HIV-1 p24 antigen achieved a detection limit of 1.49 pg/ml (He Q. et al., 2014). Detection of AuNPs has also served as a proxy for assays of anti-erythropoietin antibodies (Lu et al., 2009), virus-specific RNA (Hsu et al., 2011), human vascular endothelial growth factor (Thompson et al., 2010), and sialic acids on cancer cells (Zhang et al., 2016). Use of AuNPs for PfHRP-II biomolecular analysis has also been described (Wilschefski and Baxter, 2019). Because of its unmatched sensitivity for a wide range of metals and several non-metals, ICP-MS has increasingly been used for bioanalytic quantifications with element-tagged immunoassays in clinical diagnosis and single-cell analysis (Liu et al., 2014; Wilschefski and Baxter, 2019). Here, we describe the ICP-MS detection of AuNPs as a proxy for the sensitive immunoassay quantification of PfLDH, both as a purified recombinant protein and as antigen present in *P. falciparum*-infected erythrocytes.

METHODS

Plasmodium falciparum Parasite Cultures and Test Sample Preparations

P. falciparum parasites (3D7 line) were propagated *in vitro* in O^+ human red blood cells at 2% hematocrit in RPMI1640 medium with L-Glutamine, 25 mmol/L Hepes, 50 μg/ml Hypoxanthine (KD Medical, Columbia, MD) supplemented with 10 μg/ml gentamycin (Gibco, ThermoFisher Scientific, Grand Island, NY) and 0.5% Albumax I (ThermoFisher Scientific, Grand Island, NY) (Cranmer et al., 1997). Synchronization of parasite cultures with 5% sorbitol (Sigma) was performed as previously described (Lambros and Vanderberg, 1979). Parasitemias were determined from Giemsa-stained thin blood films, and erythrocytes per unit volume were counted by hemocytometer. The total number of parasite-infected erythrocytes was calculated as: (% parasitemia/100) × (erythrocytes/ml) × (ml of culture).

For sample preparations, *P. falciparum* infected erythrocytes were washed three times with PBS (10 mmol/L PO_4^{3-}, 137 mmol/L NaCl, 2.7 mmol/L KCl) and lysed in a radioimmunoprecipitation assay buffer (RIPA buffer, Thermo Fisher Scientific, CA, USA) containing 1× protease inhibitor (Thermo Scientific Halt Protease Inhibitor Cocktail, Thermo Fisher Scientific, CA, USA). The samples were assayed immediately or aliquoted and stored at −20°C.

AuNP-Labeled PfLDH Antibodies and Sample Preparations

Malaria PfLDH mouse monoclonal antibodies (mAbs) MBS498007 and MBS498008 were purchased from MyBioSource, Inc. (San Diego, CA). Detection mAb MBS498008 was labeled with AuNPs using a 60 nm gold conjugation kit (ab188216, Abcam, CA), which covalently attached the ultra-stable AuNPs. Briefly, stock mAb MBS498008 (5 mg/ml) was diluted to 0.1 mg/ml with the gold 60 nm Antibody Diluent. For each reaction, 42 µl of the 60 nm AuNP reaction buffer was added to 12 µl of diluted antibody and mixed thoroughly by pipetting up and down at least 5 times. Forty-five µl of the mixture was then transferred to a vial of 60 nm AuNPs (provided in the Abcam kit) and mixed by gently pipetting up and down. Finally, 5 µl of quencher reagent was added and the AuNP-labeled mAb was stored at 4°C before use.

Immunoassay Detection of PfLDH Antigens

The PfLDH immunoassay was performed as previously described (Mu et al., 2017) with some modifications. Test tubes (TBS0211, Bio-Rad) were coated with anti-PfLDH mAb MBS498007 (6 µg/ml) overnight at 4°C. After three washes with PBS plus 0.05% Tween 20, the plates were blocked with tris-buffered saline, 4% bovine serum albumin (BSA), 0.05% Tween 20 (MilliporeSigma) for 2 h. Two-fold serial dilutions of recombinant PfLDH protein (MBS5308810, MyBioSource, Inc.) in 30 µl PBS buffer or 3D7 parasite lysate in RIPA buffer were added and incubated for 2 h at 37°C. After six additional washes, 30 µl of AuNP-labeled mAb MBS498008 (200 ng/ml) was incubated with the captured antigen for 1 h at 37°C. Unlabeled mAb was removed by five washes and samples were stored dry in the test tubes at 4°C before the ICP-MS was performed.

Inductively Coupled Plasma Mass Spectrometry

ICP-MS was performed as described previously (Assumpção et al., 2013). Briefly, a 200 µl aliquot of freshly prepared aqua regia, a mixture of one part nitric acid and three parts hydrochloric acid by volume, was deposited into the test tube. The tubes were capped, and the contents were allowed to react at room temperature of 21°C on the bench overnight. The contents in the tube were quantitatively transferred to a pre-weighed 15 ml Falcon centrifuge tube (Corning 352097, Thermo Fisher Scientific) and diluted to 5 ml with locally purified water distilled below the boiling point. The mass of the resulting sample was weighed on a Mettler Toledo (Columbus, OH) model AT261 Delta Range analytical balance by difference. Procedural blanks were prepared similarly. Samples and the blanks were analyzed using the SemiQuant mode of Agilent Technologies (Santa Clara, CA) Model 7500CS inductively coupled plasma mass spectrometer. The sample introduction system of the mass spectrometer consisted of a perfluoroalkoxy alkanes (PFA) microflow (0.1 ml/min) concentric nebulizer, a PFA Scott-type double-pass spray chamber, and a sapphire injector. The SemiQuant mode is capable of quantifying elements of the entire periodic table by using the spectral information of the elements. The instrument was calibrated using a solution containing 20 µg/kg each of 31 elements (lithium, beryllium, boron, sodium, magnesium, aluminum, calcium, scandium, vanadium, chromium, manganese, cobalt, nickel, copper, zinc, arsenic, selenium, strontium, molybdenum, silver, cadmium, antimony, barium, lanthanum, europium, holmium, ytterbium, thallium, lead, thorium, and uranium) prepared by diluting Catalog No. ICP-MSCS ICP-MS Calibration Standard (High Purity Standards, Charleston, SC) with 1.5% volume fraction nitric acid (Optima grade, Thermo Fisher Scientific) in water. National Institute of Standards and Technology Standard Reference Material (SRM) 1643f, Trace Elements in Water, was measured with the samples, serving as the quality assurance for the measurement.

Determinations of PfLDH-Based Immunoassay Sensitivity Limits

PfLDH determinations of LOD were calculated from serially-diluted samples of recombinant PfLDH protein (10 pg/ml to 0 pg/ml) by the formula: LOD = 3.3(Sy/S), where Sy and S represent the standard deviation of the response (Sy), and slope of the calibration curve (S), respectively (Evard et al., 2016). S and Sy values were obtained using "SLOPE" and "STEYX" functions in MS Excel. *P. falciparum* erythrocytic stage detection sensitivities were determined by two-fold serial-dilutions of parasite-infected cultures from 40 p/µl to 0.07 p/µl. Synchronized ring-stage and mixed stage parasite cultures were evaluated separately. ICP-MS background signals from uninfected control erythrocyte dilutions (0 p/µl) were subtracted against signals from the parasite-infected cell dilutions for analysis. One-way ANOVA and 95% confidence intervals were calculated in Prism 8 (GraphPad Software, La Jolla, CA, USA). All experiments were performed in duplicate.

RESULTS

Sensitive Immunoassay Detection of PfLDH by Inductively Coupled Plasma Mass Spectrometry of Bound AuNP-Labeled Antibody

For the immunoassay detection of PfLDH, capture of the protein by the primary capture mAb and binding of the detection mAb were as previously described for high sensitivity immuno-PCR assays (Mu et al., 2017). However, instead of the biotinylated DNA tag used for immuno-PCR, the detection mAb was labeled with AuNPs (**Figure 1**). Following disruption of the immune complex with aqua regia solution, then vaporization and

ionization in the nebulizer and spray chamber, gold atoms were separated according to their mass-charge ration (m/z) and quantified by ICP-MS. The amount of PfLDH in the sample is calculated with the following equation:

$$C_{PfLDH_sample} = \frac{I_{Au_sample} * C_{PfLDH_standard}}{I_{Au_PfLDH_standard}} \quad (1)$$

where C_{PfLDH_sample} and $C_{fLDH_standard}$ are the concentration of PfLDH in the sample and the standard, respectively; I_{Au_sample} and $I_{Au_PfLDH_standard}$ are the instrument response to ^{197}Au in the PfLDH sample and the standard, respectively. To calibrate detection sensitivity, 30 µl volumes of a dilution series of recombinant PfLDH protein dilutions from 0.3 pg/ml to 10 pg/ml were tested by the assay protocol. ICP-MS quantification of the bound gold atoms (**Figure 2**) demonstrated sensitive detection of PfLDH protein in the pg/ml range. Note that each point on the curve was obtained from a PfLDH sample subjected to the AuNP tagged immunoassay procedure. The results from the negative controls (**Table 1**) showed that the carry-over of AuNP or unbound AuNP reagents had been effectively minimized by the multiple washes with PBS. Linear fit of the data yielded a calibration curve expressed as $y = 2.370x - 0.772$, $R^2 = 0.9839$, with a PfLDH LOD of 1.5 pg/ml.

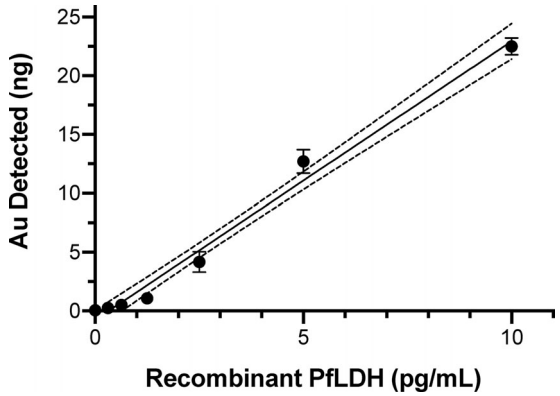


FIGURE 2 | ICP-MS immunoassay results showing the elemental gold signal levels from known amounts of recombinant PfLDH protein. PfLDH was captured by the primary mAb from 30 µl volumes of the indicated concentrations of PfLDH. After binding of AuNP-labeled detection mAb and analysis by ICP-MS, the elemental gold levels were used to calculate the linear best-fit calibration line (solid line) and 95% confidence intervals (dotted lines). Average values with standard errors are shown. All experiments were performed in duplicate.

Sensitive Detection of *Plasmodium falciparum*-Infected Erythrocytes by the Inductively Coupled Plasma Mass Spectrometry PfLDH Immunoassay

We next evaluated the sensitivity of the ICP-MS PfLDH immunoassay for detection of *P. falciparum*-infected erythrocytes. Experiments were performed with mixed stage-infected erythrocytes (**Figure 3A**) or with synchronized ring stage-infected erythrocytes (**Figure 3B**). The Au signals at 0.57 to 0.60 ng from the lowest test parasitemia (0.07 p/µl) were more than 10× higher compared to that of the control sample (0.05 ng, **Table 1**), testifying to the high sensitivity of ICP-MS based immunoassay in detection of *P. falciparum* infections. Statistical analysis with 95% confidence interval of difference (**Table 1**), showed that the significant difference between the test samples and control could be confidently detected down to a level of 0.32 p/µl for mixed stage (**Figure 3C**) or 1.6 p/µl for ring stage test samples (**Figure 3D**).

DISCUSSION

Recent years have seen remarkable advancement in the application of ICP-MS technology for molecular diagnostics,

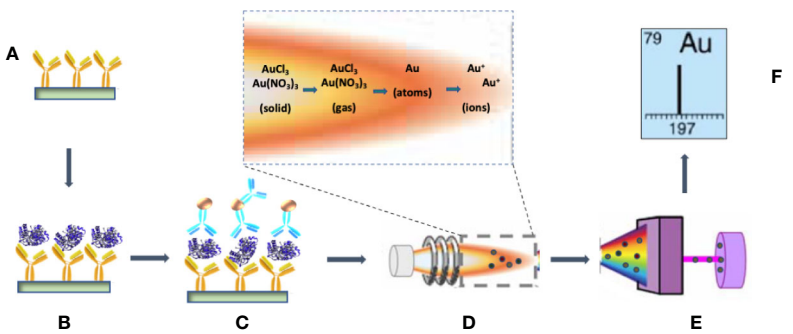


FIGURE 1 | Schematic of the processes involved in the immunoassay detection of PfLDH by ICP-MS. **(A)** The test tubes are coated with the primary anti-PfLDH monoclonal antibody (mAb). **(B)** After blocking with BSA and washing, PfLDH antigens are captured by the primary mAb. **(C)** AuNP-labeled detection mAbs are bound to the capture PfLDH molecules. **(D)** The AuNP–protein complexes are dissolved in the aqua regia for injection into the spray chamber. ICP, which is a plasma sustained by inductively coupled radio frequency energy *via* a load coil after generation from a stream of argon passed through a quartz torch, serves as the ionization source. In the vaporization and ionization phase, the droplets of the sample travel through the different heating zones, in which they are dried by the plasma torch, vaporized, atomized, and finally ionized. **(E)** Ions are separated by the mass analyzer. **(F)** The detector displays the isotopic fingerprint and the analyte ion count rate, which is proportional to the mass fraction of the analyte in the sample droplets.

TABLE 1 | AuNP quantifications of parasite-infected relative to control erythrocyte samples by ICP-MS immunoassays.[a]

	Sample (p/μl)	AuNP quantification (ng)[c]	PfLDH amount (pg/ml)	PfLDH Mean Difference - Parasites *vs* control (pg/ml)	95% CI [b] (pg/ml)	Adjusted P Value	Significance
Mixed stage parasitemia	0.07	0.60	0.58	0.56	–0.11 to 1.23	0.671	ns
	0.32	2.21	1.26	1.24	0.56 to 1.91	0.0133	*
	1.6	3.31	1.92	1.70	0.78 to 2.13	0.0046	**
	8	4.72	2.32	2.30	1.62 to 2.97	0.0002	***
	40	13.3	5.92	5.90	5.22 to 6.57	<0.0001	****
Control	0	0.05					
Ring stage parasitemia	0.07	0.57	0.57	0.55	–0.51 to 1.60	0.9562	ns
	0.32	1.30	0.87	0.85	–0.20 to 1.90	0.4836	ns
	1.6	2.73	1.48	1.46	0.65 to 2.75	0.0101	*
	8	4.97	2.42	2.40	1.36 to 3.45	0.0004	***
	40	9.56	4.36	4.34	3.29 to 5.39	<0.0001	****
Control	0	0.05					

[a]*One-way ANOVA analysis was used for multiple comparisons test.*

[b]*95% confidence interval of difference. Significance indications: ns, not significant; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001.*

[c]*Mean of two independent biological repeats.*

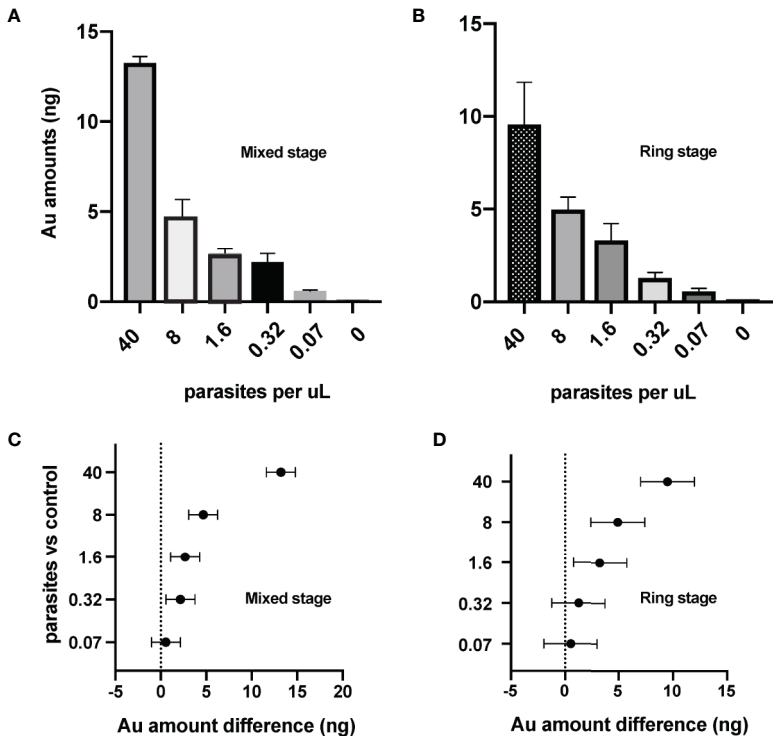


FIGURE 3 | Sensitive detection of the *P. falciparum* parasites by ICP-MS immunoassay targeting the PfLDH antigen. The ICP-MS method is able to detect AuNPs from different samples of **(A)** mixed stage and **(B)** ring stage *P. falciparum*-infected erythrocytes. Limits of detection (LOD) of the ICP-MS PfLDH immunoassay are estimated to be **(C)** 0.3 p/μl for the mixed stages, and **(D)** 1.6 p/μl for the ring stages. The horizontal bars represent the 95% confidence interval. The averages are from experiments repeated separately on different days.

particularly when it is applied to molecular targets with metal nanoparticle labels (Wilschefski and Baxter, 2019). Adding to the growing list, the ICP-MS PfLDH immunoassay described here detects *P. falciparum*-infected erythrocytes at the level of ≈1 p/μl, comparable to the detection limit of immuno-PCR in sandwich assays with the same mAbs (Mu et al., 2017).

A mass spectrometer is an ion counter. By labeling PfLDH with AuNP, the analyte density as defined by the number of

PfLDH molecules is effectively multiplied by a factor which is the number of Au atoms in the average number of AuNPs attached to each PfLDH molecule. Note that the number of Au atoms, and hence the signal enhancement factor by ICP-MS detection, is proportional to the third power of the diameter of attached AuNPs. It may therefore be possible to improve detection sensitivity further by using AuNPs of larger diameter for labeling. With the 60 nm AuNP, ICP-MS is already capable of detecting the amount of Au at 10× below the amount associated with 0.07 p/μl, which is far below the detection limit of 0.32 p/μl to 1.6 p/μl discussed above. Therefore, the limiting factor to the LOD of the parasite is the reproducibility of the sandwich immunoassay procedure especially at low PfLDH concentration end rather than ICP-MS detection of Au. Strategies such as chemically modified nucleotide-based elemental tags (Hu et al., 2019), three-dimensional structure modifications of the AuNPs (Wang et al., 2018), rolling circle amplification of parasite DNA detected with AuNP-sensors (He Y. et al., 2014), and multiplex assays of biomarker proteins with AuNP-tagged antibodies (Terenghi et al., 2009) may also provide opportunities for higher sensitivity detection by ICP-MS. The ultimate goal will be to develop high-sensitivity detection of AuNPs with practical, field-robust devices at point of care. Design and development of the necessary equipment, the possibility of which brings to mind the successful production of a low-cost, portable, field-programmable nuclear magnetic resonance spectrometer (Peng et al., 2014), will support this goal.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JM and TW initiated the study and drafted the manuscript. JM and LY conducted the lab work and analyzed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Intramural Research Program of the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health.

ACKNOWLEDGMENTS

We express our thanks to Roy Faiman, Kristin D. Lane, T. Parks Remcho, and Juliana M. Sá for discussions and comments on the manuscript.

REFERENCES

Assumpção, T. C., Ma, D., Schwarz, A., Reiter, K., Santana, J. M., Andersen, J. F., et al. (2013). Salivary antigen-5/CAP family members are Cu2+-dependent antioxidant enzymes that scavenge O_{2}-. and inhibit collagen-induced platelet aggregation and neutrophil oxidative burst. *J. Biol. Chem.* 288, 14341–14361. doi: 10.1074/jbc.M113.466995

Bell, D., Wongsrichanalai, C., and Barnwell, J. W. (2006). Ensuring quality and access for malaria diagnosis: how can it be achieved? *Nat. Rev. Microbiol.* 4, S7–20. doi: 10.1038/nrmicro1525

Bosco, A. B., Anderson, K., Gresty, K., Prosser, C., Smith, D., Nankabirwa, J. I., et al. (2020). Molecular surveillance reveals the presence of pfhrp2 and pfhrp3 gene deletions in Plasmodium falciparum parasite populations in Uganda 2017-2019. *Malar J.* 19, 300. doi: 10.1186/s12936-020-03362-x

Bourgeois, N., Boutet, A., Bousquet, P. J., Basset, D., Douard-Enault, C., Charachon, S., et al. (2010). Comparison of three real-time PCR methods with blood smears and rapid diagnostic test in Plasmodium sp. infection. *Clin. Microbiol. Infect.* 16, 1305–1311. doi: 10.1111/j.1469-0691.2009.02933.x

Castilho Mde, S., Laube, T., Yamanaka, H., Alegret, S., and Pividori, M. I. (2011). Magneto immunoassays for Plasmodium falciparum histidine-rich protein 2 related to malaria based on magnetic nanoparticles. *Anal. Chem.* 83, 5570–5577. doi: 10.1021/ac200573s

Chakma, B., Jain, P., Singh, N. K., and Goswami, P. (2016). Development of an indicator displacement based detection of malaria targeting HRP-II as biomarker for application in point-of-care settings. *Anal. Chem.* 88, 10316–10321. doi: 10.1021/acs.analchem.6b03315

Cordeiro, M., Ferreira Carlos, F., Pedrosa, P., Lopez, A., and Baptista, P. V. (2016). Gold nanoparticles for diagnostics: advances towards points of care. *Diagnostics (Basel)* 6, 43. doi: 10.3390/diagnostics6040043

Cranmer, S. L., Magowan, C., Liang, J., Coppel, R. L., and Cooke, B. M. (1997). An alternative to serum for cultivation of Plasmodium falciparum in vitro. *Trans. R Soc. Trop. Med. Hyg.* 91, 363–365. doi: 10.1016/S0035-9203(97)90110-3

Cunningham, C. H., Hennelly, C. M., Lin, J. T., Ubalee, R., Boyce, R. M., Mulogo, E. M., et al. (2020). A novel CRISPR-based malaria diagnostic capable of Plasmodium detection, speciation, and drug-resistance genotyping. *bioRxiv*. 2020.2004.2001.017962. doi: 10.1101/2020.04.01.017962

Das, S., Peck, R. B., Barney, R., Jang, I. K., Kahn, M., Zhu, M., et al. (2018). Performance of an ultra-sensitive Plasmodium falciparum HRP2-based rapid diagnostic test with recombinant HRP2, culture parasites, and archived whole blood samples. *Malar J.* 17, 118. doi: 10.1186/s12936-018-2268-7

Dirkzwager, R. M., Liang, S., and Tanner, J. A. (2016). Development of aptamer-based point-of-care diagnostic devices for malaria using three-dimensional printing rapid prototyping. *ACS Sensors* 1, 420–426. doi: 10.1021/acssensors.5b00175

Evard, H., Kruve, A., and Leito, I. (2016). Tutorial on estimating the limit of detection using LC-MS analysis, part II: Practical aspects. *Anal. Chim. Acta* 942, 40–49. doi: 10.1016/j.aca.2016.08.042

Fraser, L. A., Kinghorn, A. B., Dirkzwager, R. M., Liang, S., Cheung, Y. W., Lim, B., et al. (2018). A portable microfluidic Aptamer-Tethered Enzyme Capture (APTEC) biosensor for malaria diagnosis. *Biosens. Bioelectron* 100, 591–596. doi: 10.1016/j.bios.2017.10.001

Gatton, M. L., Dunn, J., Chaudhry, A., Ciketic, S., Cunningham, J., and Cheng, Q. (2017). Implications of parasites lacking Plasmodium falciparum histidine-rich protein 2 on malaria morbidity and control when rapid diagnostic tests are used for diagnosis. *J. Infect. Dis.* 215, 1156–1166. doi: 10.1093/infdis/jix094

Gitta, B., and Kilian, N. (2020). Diagnosis of malaria parasites plasmodium spp. in endemic areas: current strategies for an ancient disease. *Bioessays* 42, e1900138. doi: 10.1002/bies.201900138

Grignard, L., Nolder, D., Sepulveda, N., Berhane, A., Mihreteab, S., Kaaya, R., et al. (2020). A novel multiplex qPCR assay for detection of Plasmodium falciparum with histidine-rich protein 2 and 3 (pfhrp2 and pfhrp3) deletions in polyclonal infections. *EBioMedicine* 55, 102757. doi: 10.1016/j.ebiom.2020.102757

Guirgis, B. S., Sa E Cunha, C., Gomes, I., Cavadas, M., Silva, I., Doria, G., et al. (2012). Gold nanoparticle-based fluorescence immunoassay for malaria antigen detection. *Anal. Bioanal. Chem.* 402, 1019–1027. doi: 10.1007/s00216-011-5489-y

Han, E. T., Watanabe, R., Sattabongkot, J., Khuntirat, B., Sirichaisinthop, J., Iriko, H., et al. (2007). Detection of four Plasmodium species by genus- and species-specific loop-mediated isothermal amplification for clinical diagnosis. *J. Clin. Microbiol.* 45, 2521–2528. doi: 10.1128/JCM.02117-06

He, Q., Zhu, Z. L., Jin, L. L., Peng, L., Guo, W., and Hu, S. H. (2014). Detection of HIV-1 p24 antigen using streptavidin-biotin and gold nanoparticles based immunoassay by inductively coupled plasma mass spectrometry. *J. Anal. Atomic Spectrom.* 29, 1477–1482. doi: 10.1039/C4JA00026A

He, Y., Chen, D., Li, M., Fang, L., Yang, W., Xu, L., et al. (2014). Rolling circle amplification combined with gold nanoparticles-tag for ultra sensitive and specific quantification of DNA by inductively coupled plasma mass spectrometry. *Biosens. Bioelectron* 58, 209–213. doi: 10.1016/j.bios.2014.02.072

Hemben, A., Ashley, J., and Tothill, I. E. (2017). Development of an immunosensor for PfHRP 2 as a biomarker for malaria detection. *Biosens. (Basel)* 7, 28. doi: 10.3390/bios7030028

Hemben, A., Ashley, J., and Tothill, I. E. (2018). An immunosensor for parasite lactate dehydrogenase detection as a malaria biomarker - comparison with commercial test kit. *Talanta* 187, 321–329. doi: 10.1016/j.talanta.2018.04.086

Hsu, I. H., Chen, W. H., Wu, T. K., and Sun, Y. C. (2011). Gold nanoparticle-based inductively coupled plasma mass spectrometry amplification and magnetic separation for the sensitive detection of a virus-specific RNA sequence. *J. Chromatogr. A* 1218, 1795–1801. doi: 10.1016/j.chroma.2011.02.005

Hu, Z., Sun, G., Jiang, W., Xu, F., Zhang, Y., Xia, M., et al. (2019). Chemical-modified nucleotide-based elemental tags for high-sensitive immunoassay. *Anal. Chem.* 91, 5980–5986. doi: 10.1021/acs.analchem.9b00405

Jeon, W., Lee, S., Manjunatha, D. H., and Ban, C. (2013). A colorimetric aptasensor for the diagnosis of malaria based on cationic polymers and gold nanoparticles. *Anal. Biochem.* 439, 11–16. doi: 10.1016/j.ab.2013.03.032

Jiang, P., Wang, Y., Zhao, L., Ji, C., Chen, D., and Nie, L. (2018). Applications of gold nanoparticles in non-optical biosensors. *Nanomaterials (Basel)* 8, 977. doi: 10.3390/nano8120977

Kongkasuriyachai, D., Yongkiettrakul, S., Kiatpathomchai, W., and Arunrut, N. (2017). Loop-mediated isothermal amplification and LFD combination for detection of Plasmodium falciparum and Plasmodium vivax. *Methods Mol. Biol.* 1572, 431–443. doi: 10.1007/978-1-4939-6911-1_28

Krampa, F. D., Aniweh, Y., Kanyong, P., and Awandare, G. A. (2020). Recent advances in the development of biosensors for malaria diagnosis. *Sensors (Basel)* 20, 799. doi: 10.3390/s20030799

Lambros, C., and Vanderberg, J. P. (1979). Synchronization of Plasmodium falciparum erythrocytic stages in culture. *J. Parasitol.* 65, 418–420. doi: 10.2307/3280287

Lee, S., Song, K. M., Jeon, W., Jo, H., Shim, Y. B., and Ban, C. (2012). A highly sensitive aptasensor towards Plasmodium lactate dehydrogenase for the diagnosis of malaria. *Biosens. Bioelectron* 35, 291–296. doi: 10.1016/j.bios.2012.03.003

Liu, R., Wu, P., Yang, L., Hou, X., and Lv, Y. (2014). Inductively coupled plasma mass spectrometry-based immunoassay: a review. *Mass Spectrom. Rev.* 33, 373–393. doi: 10.1002/mas.21391

Low, Y. K., Chan, J., Soraya, G. V., Buffet, C., Abeyrathne, C. D., Huynh, D. H., et al. (2019). Development of an ultrasensitive impedimetric immunosensor platform for detection of Plasmodium lactate dehydrogenase. *Sensors (Basel)* 19, 2446. doi: 10.3390/s19112446

Lu, Y., Wang, W., Xing, Z., Wang, S., Cao, P., Zhang, S., et al. (2009). Development of an ICP-MS immunoassay for the detection of anti-erythropoietin antibodies. *Talanta* 78, 869–873. doi: 10.1016/j.talanta.2008.12.065

Mohon, A. N., Getie, S., Jahan, N., Alam, M. S., and Pillai, D. R. (2019). Ultrasensitive loop mediated isothermal amplification (US-LAMP) to detect malaria for elimination. *Malar J.* 18, 350. doi: 10.1186/s12936-019-2979-4

Mu, J., Andersen, J. F., Valenzuela, J. G., and Wellems, T. E. (2017). High-sensitivity assays for Plasmodium falciparum infection by immuno-polymerase chain reaction detection of PfIDEh and PfLDH antigens. *J. Infect. Dis.* 216, 713–722. doi: 10.1093/infdis/jix369

Obisesan, O. R., Adekunle, A. S., Oyekunle, J., Sabu, T., Nkambule, T. T. I., and Mamba, B. B. (2019). Development of electrochemical nanosensor for the detection of malaria parasite in clinical samples. *Front. Chem.* 7, 89. doi: 10.3389/fchem.2019.00089

Peng, W. K., Kong, T. F., Ng, C. S., Chen, L., Huang, Y., Bhagat, A. A., et al. (2014). Micromagnetic resonance relaxometry for rapid label-free malaria diagnosis. *Nat. Med.* 20, 1069–1073. doi: 10.1038/nm.3622

Poschl, B., Waneesorn, J., Thekisoe, O., Chutipongvivate, S., and Karanis, P. (2010). Comparative diagnosis of malaria infections by microscopy, nested PCR, and LAMP in northern Thailand. *Am. J. Trop. Med. Hyg.* 83, 56–60. doi: 10.4269/ajtmh.2010.09-0630

Rifaie-Graham, O., Pollard, J., Raccio, S., Balog, S., Rusch, S., Hernandez-Castaneda, M. A., et al. (2019). Hemozoin-catalyzed precipitation polymerization as an assay for malaria diagnosis. *Nat. Commun.* 10, 1369. doi: 10.1038/s41467-019-09122-z

Sharma, M. K., Rao, V. K., Merwyn, S., Agarwal, G. S., Upadhyay, S., and Vijayaraghavan, R. (2011). A novel piezoelectric immunosensor for the detection of malarial Plasmodium falciparum histidine rich protein-2 antigen. *Talanta* 85, 1812–1817. doi: 10.1016/j.talanta.2011.07.008

Slater, H. C., Ross, A., Ouedraogo, A. L., White, L. J., Nguon, C., Walker, P. G., et al. (2015). Assessing the impact of next-generation rapid diagnostic tests on Plasmodium falciparum malaria elimination strategies. *Nature* 528, S94–101. doi: 10.1038/nature16040

Terenghi, M., Elviri, L., Careri, M., Mangia, A., and Lobinski, R. (2009). Multiplexed determination of protein biomarkers using metal-tagged antibodies and size exclusion chromatography–inductively coupled plasma mass spectrometry. *Anal. Chem.* 81, 9440–9448. doi: 10.1021/ac901853g

Thompson, D. F., Eborall, W., Dinsmore, A., Smith, C. J., and Duckett, C. J. (2010). Development and validation of a NANOGold immunoassay for the detection of vascular endothelial growth factor (VEGF) in human serum using inductively coupled plasma mass spectrometry. *Rapid Commun. Mass Spectrom.* 24, 927–932. doi: 10.1002/rcm.4456

Verma, A. K., Bharti, P. K., and Das, A. (2018). HRP-2 deletion: a hole in the ship of malaria elimination. *Lancet Infect. Dis.* 18, 826–827. doi: 10.1016/S1473-3099(18)30420-1

Wang, X., Du, D., Dong, H., Song, S., Koh, K., and Chen, H. (2018). para-Sulfonatocalix[4]arene stabilized gold nanoparticles multilayers interfaced to electrodes through host-guest interaction for sensitive ErbB2 detection. *Biosens. Bioelectron* 99, 375–381. doi: 10.1016/j.bios.2017.08.011

Wilschefski, S. C., and Baxter, M. R. (2019). Inductively coupled plasma mass spectrometry: introduction to analytical aspects. *Clin. Biochem. Rev.* 40, 115–133. doi: 10.33176/AACB-19-00024

Wu, L., Van Den Hoogen, L. L., Slater, H., Walker, P. G., Ghani, A. C., Drakeley, C. J., et al. (2015). Comparison of diagnostics for the detection of asymptomatic Plasmodium falciparum infections to inform control and elimination strategies. *Nature* 528, S86–S93. doi: 10.1038/nature16039

Zhang, X., Chen, B., He, M., Zhang, Y., Peng, L., and Hu, B. (2016). Boronic acid recognition based-gold nanoparticle-labeling strategy for the assay of sialic acid expression on cancer cell surface by inductively coupled plasma mass spectrometry. *Analyst* 141, 1286–1293. doi: 10.1039/C5AN02402A

Disclaimer: Certain commercial equipment, instruments or materials are identified in this manuscript to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

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.

Copyright © 2021 Mu, Yu and Wellems. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

REVIEW
published: 21 January 2021
doi: 10.3389/fmicb.2020.612568

An Evolutionary Arms Race Between *Burkholderia pseudomallei* and Host Immune System: What Do We Know?

Chalita Chomkatekaew[1†], Phumrapee Boonklang[1†], Apiwat Sangphukieo[1,2] and Claire Chewapreecha[1,2,3]*

[1] Mahidol-Oxford Tropical Medicine Research Unit (MORU), Bangkok, Thailand, [2] Bioinformatics and Systems Biology Program, School of Bioresource and Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand, [3] Wellcome Sanger Institute, Hinxton, United Kingdom

OPEN ACCESS

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

Reviewed by:
Apichai Tuanyok,
University of Florida, United States
Sheila Nathan,
National University of Malaysia, Malaysia
Yunn Hwen Gan,
National University of Singapore, Singapore

***Correspondence:**
Claire Chewapreecha
claire@tropmedres.ac

[†] 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: 30 September 2020
Accepted: 21 December 2020
Published: 21 January 2021

Citation:
Chomkatekaew C, Boonklang P, Sangphukieo A and Chewapreecha C (2021) An Evolutionary Arms Race Between Burkholderia pseudomallei and Host Immune System: What Do We Know?
Front. Microbiol. 11:612568.
doi: 10.3389/fmicb.2020.612568

A better understanding of co-evolution between pathogens and hosts holds promise for better prevention and control strategies. This review will explore the interactions between *Burkholderia pseudomallei,* an environmental and opportunistic pathogen, and the human host immune system. *B. pseudomallei* causes "Melioidosis," a rapidly fatal tropical infectious disease predicted to affect 165,000 cases annually worldwide, of which 89,000 are fatal. Genetic heterogeneities were reported in both *B. pseudomallei* and human host population, some of which may, at least in part, contribute to inter-individual differences in disease susceptibility. Here, we review (i) a multi-host—pathogen characteristic of the interaction; (ii) selection pressures acting on *B. pseudomallei* and human genomes with the former being driven by bacterial adaptation across ranges of ecological niches while the latter are driven by human encounter of broad ranges of pathogens; (iii) the mechanisms that generate genetic diversity in bacterial and host population particularly in sequences encoding proteins functioning in host—pathogen interaction; (iv) reported genetic and structural variations of proteins or molecules observed in *B. pseudomallei*—human host interactions and their implications in infection outcomes. Together, these predict bacterial and host evolutionary trajectory which continues to generate genetic diversity in bacterium and operates host immune selection at the molecular level.

Keywords: melioidosis, evolution, burkholderia, host immune sytem, genetic variants

INTRODUCTION

Melioidosis is a serious and often fatal neglected tropical infectious disease caused by *Burkholderia pseudomallei*, an intracellular bacterial pathogen and also ubiquitous in the environment (Holden et al., 2004). Human hosts can acquire the bacterium after direct environmental exposure either through dermal puncture, ingestion of contaminated food or water supplies, or inhalation of contaminated soil or water aerosols. Following an acquisition, *B. pseudomallei* can replicate in host non-phagocytic and phagocytic cells as well as spreading intracellularly (Willcocks et al., 2016). The bacterium can either kill the human host rapidly (acute melioidosis) or hide within the host body for a long period (chronic melioidosis). With rare exceptions (Abbink et al., 2001; Ralph et al., 2004; Aziz et al., 2020), *B. pseudomallei* is not known to transmit from person to person; indicating

that *B. pseudomallei* has not evolved virulence mechanisms through consecutive passage in human hosts. Moreover, *B. pseudomallei* within-host evolution observed in chronic infections are often linked to attenuated virulence (Price et al., 2013; Viberg et al., 2017; Pearson et al., 2020), which is possibly mediated by host immune evasion. While environment exposure can lead to melioidosis, most patients are elderly or have one or more underlying health risks. The most common health condition is diabetes mellitus which presents in up to half of cases (Kronsteiner et al., 2019). Therefore, it is likely that *B. pseudomallei* has acquired virulence genes for human infection before being exposed to the human hosts, and that these virulence factors are most effective when the host has underlying health issues. It is possible that a successful human infection is mediated by bacterial redundant virulence mechanisms that were acquired and maintained in the environment. In this review, we will leverage recent pieces of genomic studies to better understand the evolution of virulence mechanisms in *B. pseudomallei* and host defense system.

A MULTI-HOST—PATHOGEN INTERACTION

Interactions between free-living amoebae and environmental bacteria have been proposed to provide an evolutionary training platform for intracellular pathogen such as *Legionella pnemophila* (Park et al., 2020), *Listeria monocytogenes* (Schuppler, 2014), and *B. pseudomallei* (Inglis et al., 2000, 2003; Noinarin et al., 2016). Many free-living amoebae including the genera *Acanthamoeba*, *Dictyostelium*, *Naegleria*, and *Paravahlkampfia* share natural habitats with *Burkholderia* bacteria, including *B. pseudomallei* (Inglis et al., 2000, 2003; Noinarin et al., 2016). Amoeba-*Burkholderia* interactions can range from predator-prey to mutualistic relationships. The fate of these interactions is subjected to the species of amoebae host and *Burkholderia* (genotypes), stages of host, as well as external factors that modify the outcomes. The social amoebae *Dictyostelium discoideum* life cycle ranges from unicellular amoebae to multicellular slugs where ten of thousands of single-celled *D. discoideum* aggregate to form a fruiting body. *D. discoideum* can directly ingest *B. pseudomallei* and other *Burkholderia* bacteria as their food; thereby requiring the bacterium to resist host cell phagocytosis, persist inside a unicellular host, and migrate between host cells through the amoeba cytoskeleton. These processes share many similarities with mammalian host infection and likely prime the bacterium for an intracellular lifestyle. The sentinel cells in *D. discoideum*, which make up approximately 1% of slug cells, can use antimicrobial defense systems similar to those employed by phagocytes in the human innate immune system (Zhang et al., 2016). The sentinel cells are capable of releasing reactive oxygen species to lyse the soil bacterium. Moreover, they can also release an extracellular trap—the reticulated nets of DNA carrying antimicrobial granules—to kill the invading bacteria. The extracellular trap is an ancient host-defense mechanism common among phagocytic cells across vertebrates and invertebrates, thereby providing a training ground for *B. pseudomallei* infection in mammalian hosts. External factors such as nutrient availability were shown to determine the fate of *D. discoideum—Burkholderia* relationship. The association is beneficial to both parties when the nutrient is scarce. Under nutrient limited conditions, *Burkholderia*—associated *D. discoideum* produced more spores (Brock et al., 2013; DiSalvo et al., 2015) and had an increased uptake of secondary bacteria that can be used as food (Khojandi et al., 2019) which led to a better survival rate than *D. discoideum* without *Burkholderia* association. Under this condition, *Burkholderia* could be detected inside *D. discoideum* spores, which enable the bacterium to better disseminate. The mechanism underlining the decision to kill or cooperate is unclear.

The species of *Burkholderia* and genetic variations within the species have been shown to impact the outcomes of amoeba internalization (Haselkorn et al., 2019). It is possible that genetic variations in *B. pseudomallei* could influence the outcomes of human melioidosis. Heritability scores (Finucane et al., 2015; Hou et al., 2019; Speed et al., 2020) can be used to quantify the proportion of variations in the infection outcomes that can be explained by genetic variations in pathogens and hosts. The technique has been successfully applied to different pathogens (Lees et al., 2017) and human infectious diseases. However, the concept has not been widely adopted for *B. pseudomallei* – host infection. For *B. pseudomallei* infection in human hosts with no comorbidity, we estimated that 8% of host mortality could be explained by *B. pseudomallei* genotypes ($h^2 = 0.081$, $SE = 0.050$, $p = 0.018$; see **Supplementary Text** for method). This new result highlights a moderate proportion of infection outcomes being explained by the bacterial genetics, and suggests that a substantial proportion of the outcomes could be explained by host genetics.

Interactions between pathogens and *Homo sapiens* have shaped the evolution of modern humans (Karlsson et al., 2014). Currently, less is known about how melioidosis has shaped the human populations, particularly in melioidosis endemic areas. The host immunity could function as a general defense against all invading pathogen mediated by the innate immune system, or a pathogen-specific defense facilitated through an adaptive immune response. For innate immune response, it is possible that interactions between human hosts and other common pathogens may have driven selection on host defense pathways that affect resistance. These may include human interactions with parasites or bacteria that cause malaria (Shimizu et al., 2000; Malaria Genomic Epidemiology Network, 2008; Timmann et al., 2012; Kariuki et al., 2020), tuberculosis (Mahasirimongkol et al., 2012; Thye et al., 2012), or sepsis (Southeast Asia Infectious Disease Clinical Research Network, 2017; Sweeney et al., 2018) in many tropical and subtropical countries, and cholangiocarcinoma (Zou et al., 2014) in Southeast Asia. Using h^2, previous works estimate that genetic variations in human host can explain between 32 and 52% of infectious disease susceptibilities, depending on the population studied and the infectious agents (Cooke and Hill, 2001; Jia et al., 2019). Candidate gene approaches have determined host risk- and protective genetic markers for melioidosis infection (West et al., 2012, 2013; Myers et al., 2014; Chaichana et al., 2017; Dickey et al., 2019), many of which were previously identified as gene targets for other common

pathogen infections. To cover melioidosis-specific, as well as broad-pathogen host factors, systemic and reproducible genome-wide scans are yet to be conducted.

NATURAL SELECTION IN GENES REQUIRED FOR HOST-PATHOGEN INTERACTION

Given an increasing availability of *B. pseudomallei* and host genomes, a genome-wide scan for signatures of selection, their directions (purifying, balancing, and positive selection) and their magnitudes could be very useful to investigate host-pathogen protein-protein interactions. As both host and *B. pseudomallei* migrate, detection of directional selection can be challenging after a population bottleneck. A bottleneck leads to reduced levels of genetic variation and a subsequent loss of a selection signal. For a soil microbe like *B. pseudomallei*, bacterial migration was shown to be infrequent, with major movements tracked and dated (Pearson et al., 2009; Price et al., 2016; Chewapreecha et al., 2017). The bacterial genome evolution is correlated with a strong geographical signal, with the highest diversity observed in Australia and reduced diversity following dissemination out of the Australian origin (Chewapreecha et al., 2017). Nevertheless, frequent recombination observed in *B. pseudomallei* can introduce new genetic variations, thereby enabling detection of selection signatures in other contemporary populations outside Australia (Chewapreecha et al., 2019). Natural selection on modern humans as a result of pathogen encounter and migration has been reviewed elsewhere (Karlsson et al., 2014; Sironi et al., 2015; Slodkowicz and Goldman, 2020).

For bacteria, evolutionary pressures on orthologous proteins can be quantified using the ratio between substitution rates at non-synonymous (*dN*) sites, which could have experienced selection, and synonymous (*dS*) sites which are presumably neutral. The *dN/dS* ratio is likely to be more than 1 if pressures favor changes in the protein sequence (positive selection). It is likely to be less than 1 if selections suppress changes in the protein sequence (purifying selection). Several studies (Yu et al., 2006; Losada et al., 2010; Nandi et al., 2010; Hayden et al., 2012; Chewapreecha et al., 2019) and this review utilized *dN/dS* to estimate selection pressures on coding sequences of the pathogenic *B. pseudomallei* as well as non-pathogenic but closely related species such as *B. thailandensis*. Although these studies focused on bacterial populations isolated from different melioidosis endemic regions, performed on different sample size and genetic diversity, they largely highlighted three similar patterns (**Figure 1**). The first common pattern is an elevated level of positive selection in *Burkholderia* accessory genes (genes that are variably present across different isolates) compared to core genes (genes that are consistently present in all isolates). This observation highlights the role of accessory genomes in mediating *B. pseudomallei* adaptation, a feature reflected by the open pangenome and large repertoire of accessory genes observed in *B. pseudomallei* and *B. thailandensis* (Nandi et al., 2015; Spring-Pearson et al., 2015; Chewapreecha et al., 2017). The second common pattern is the set of genes under purifying selections in *B. pseudomallei* and *B. thailandensis* involved in the bacterial replication, transcription and translation machinery; highlighting the conservation of these genes in both species. The third common pattern is that many genes under positive selection in *B. pseudomallei* and *B. thailandensis* are required for environmental survival and exchange of genetic materials including secretion systems, response regulator, heat shock proteins, and integration and/or restriction of horizontal gene transfer. However, genes required for host cell invasions such as cell adhesion, fimbriae, intracellular multiplication and macrophage killing are only positively selected in *B. pseudomallei* but not *B. thailandensis*. These distinct signals for positive selection in *B. pseudomallei* and *B. thailandensis* could mark a different pathogenic potential of the two species.

Genes under positive selection could primarily offer *B. pseudomallei* advantages to survive harsh environmental conditions. Given its tropical and sub-tropical habitats, *B. pseudomallei* experiences drought and heavy rainfall on a seasonal basis. The bacterium can be found in deeper soil layers during the dry season and moves to the soil surface during the rainy season following water movement (Thomas et al., 1979; Currie and Jacups, 2003; Manivanh et al., 2017), highlighting a spatiotemporal transition. The bacterium is also subjected to heat stress. An *in vivo* study showed that *B. pseudomallei* had a slower growth rate and a shift in gene expression toward heat-shock proteins and bacterial motility when exposed to temperature stress (Paksanont et al., 2018). Intriguingly, *B. pseudomallei* is capable of persisting in a nutrient-free distilled water (Pumpuang et al., 2011). The bacterium was subjected to distilled water since 1994 and this experiment is still ongoing, thereby marking an unusual ability of *B. pseudomallei* to survive in low nutrient media and high osmotic pressure. Lipopolysaccharide, a known virulence factor, was proposed to facilitate *B. pseudomallei* survival in water. Statistical analysis suggested that there is a correlation between the presence of LPS and rainfall, in particular, LPS serotype B (Shaw et al., 2019). An experimental study further reported that a gene of inner core LPS biosynthesis cluster (*waaE: BPSL2510)* is vital for a long-term water incubation (Moore et al., 2008). Moreover, an adhesin *BPSL1661* was identified as a hub for co-evolutionary signals in *B. pseudomallei* population (Chewapreecha et al., 2020). The gene was functionally characterized as essential during nutrient starvation (Chewapreecha et al., 2020), thereby highlighting nutrient limitation as a major evolutionary pressure experienced by this microorganism. The presence of *B. pseudomallei* is strongly associated with the low-nutrient soil (Limmathurotsakul et al., 2010b; Hantrakun et al., 2016). This led to the hypothesis that *B. pseudomallei* might have a competitive advantage over benign soil microbes in nutrient-depleted soil but is outcompeted in a nutrient-rich environment. The occurrence of *B. pseudomallei* in the nutrient-poor agricultural field has linked to crop residue burning, a common practice in the tropics where the burning practice could deplete the soil organic matter (Bot and Benites, 2005). An ability to persist in the hostile environmental conditions allows the bacterium to survive before it could seek shelters in a more protecting reservoir inside the single or multicellular hosts.

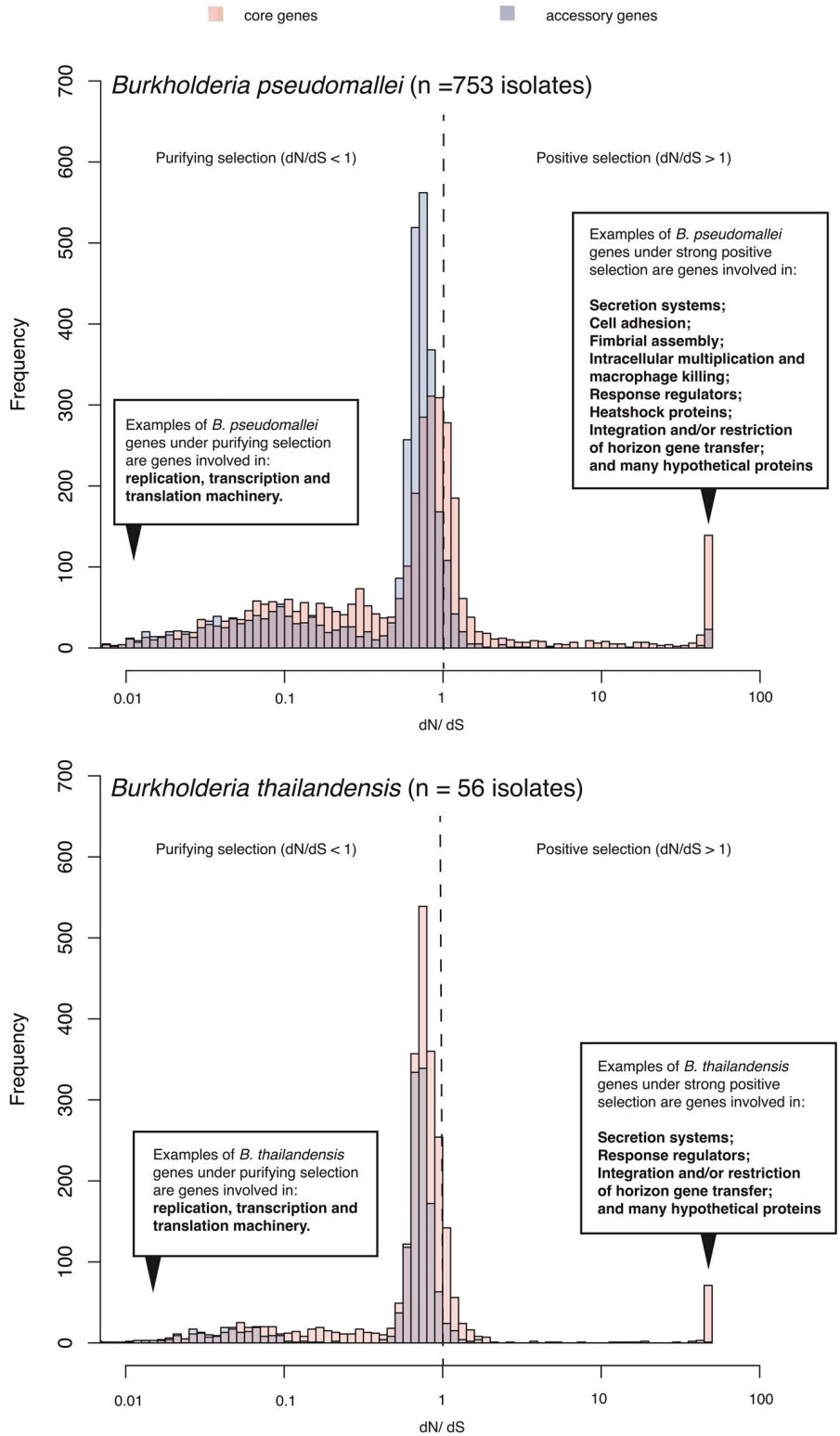


FIGURE 1 | Selection pressure acting on *B. pseudomallei* population. The histogram summarizes ranges of *dN/dS* calculated from predicted coding sequences from a collection of diverse *B. pseudomallei* population from northeast Thailand (Chewapreecha et al., 2019), and *B. thailandensis* genomes from the public database. *B. pseudomallei* and *B. thailandensis* have highly plastic genomes comprising of at least two chromosomes of approximately 7–8 Mb in size when combined. Using a pan-genome approach, all coding sequences could be categorized as "core" (present in all genomes) or "accessory" (variably present across studied genomes). Accessory genes display an elevated level of *dN/dS* which is signatures of positive selection or more relaxed purifying selection.

Genes that facilitate bacterial transition from the environment to hosts are also under positive selection. A genome-wide association study (GWAS) recently identified the bacterial genetic factors that distinguish between *B. pseudomallei* isolated from the environment and those causing disease in human (Chewapreecha et al., 2019). Genetic variants associated with cell entry and toxin were found to be more prevalent in disease- than environmental isolates, while variants involved in malfunctional cell adherence were found at higher frequency in environmental- than disease isolates. Together, this highlights significant roles of cell adhesion and cell entry in allowing the bacterium to switch to an intracellular lifestyle, either in an amoeba or in a human host where it causes melioidosis. The same study also quantified the numbers of time the variants were gained or lost from the population phylogenetic tree and highlighted multiple gain-and-loss events for both disease- and environment-associated variants. This genetic evidence suggests a process into which *B. pseudomallei* can adapt to colonize multiple niches by exploiting existing variations or exchanging disadvantageous for advantageous alleles that promote its survival in a new niche, including the human host.

MECHANISMS THAT GENERATE GENETIC AND MOLECULAR VARIATIONS

Genetic information in both *B. pseudomallei* and human host can be passed on vertically, although mistakes during replication process could result in small-scale genetic variations. In *B. pseudomallei*, larger-scale genetic variations can be introduced by recombination event or horizontal gene transfer (HGT). The former process is akin to sexual reproduction and meiotic crossing-overs in its human hosts. Due to contrasting short- and long- generation time of *B. pseudomallei* (49 min in the log phase; Ou et al., 2005) and host (22–33 years for modern humans; Ewbank, 2016); the host cannot solely rely on new genetic diversity generated when the new offspring is born. A healthy human possesses a large and dynamic repertoire of B cell- and T cell receptors that recognize the invading pathogen and produce a repertoire of antibody that recognizes a variety of antigenic structure. This molecular diversity is achieved without a change in genetic content. The mechanisms that generate the genetic and molecular diversity of the human immune system have been reviewed elsewhere (Nikolich-Zugich et al., 2004; Robinson, 2015; Dendrou et al., 2018; van den Broek et al., 2018; Adams et al., 2020).

Small-scale genetic variations in *B. pseudomallei* can be introduced by point mutation which substitutes one nucleotide with another or microindel (an insertion or deletion) that impact 1–50 bp. *B. pseudomallei* substitution rate was reported to be $1.7–4.9 \times 10^{-7}$ substitutions per site per year (Viberg et al., 2017; Pearson et al., 2020), which is comparable to those detected in other *Burkholderia* genera (Lieberman et al., 2014). The rates of indels have not been quantified in *B. pseudomallei*. However, indels can be detected from isolates obtained from acute and chronic host infections (Hayden et al., 2012; Limmathurotsakul et al., 2014a; Viberg et al., 2017) as well as from the environment (Rachlin et al., 2020), thereby highlighting the role of indels in *B. pseudomallei* evolution.

Medium-scale genetic variations can be brought in by homologous recombination, with each event contribute to a median recombining size of 5 kb (Nandi et al., 2015) (range 3 bp to 71 kb). A single recombination event can introduce 7.2 times greater nucleotide polymorphisms than a single substitution event (average r/m = 7.2). However, the amounts of SNPs being introduced per recombination event is subjected to the genetic distance between DNA donor and recipient. Sources of imported DNA could be from the same species donor, closely related species such as *B. thailandensis* and/or other soil microbes. Different *B. pseudomallei* lineages have been shown to recombine at different rates (Nandi et al., 2015), a characteristic also observed in other recombinogenic bacteria (Chewapreecha et al., 2014; Sanchez-Buso et al., 2014; David et al., 2017). Interestingly, not all DNA donor-recipient pairs are possible, suggesting a structure to the genetic flux within *B. pseudomallei* population. A lineage-specific restriction-modification system was shown to act as a barrier that restricts gene flow (Nandi et al., 2015). These modification control systems are based on DNA methylation which allows the bacterium to discriminate between correctly methylated "self" DNA, and inappropriately methylated or unmethylated "non-self" DNA. Correctly methylated DNA can be taken up and integrated into the bacterial genome, whereas inappropriately methylated or unmethylated DNA will be degraded. Interestingly, many recombination "hotspots" are focused on bacterial virulence factors (Nandi et al., 2015), highlighting the role of recombination in tuning the bacterial ability to infect hosts.

Large-scale genetic variations can be introduced by HGT or large-scale insertion or deletion. HGT includes the uptake of foreign DNA that can subsequently be integrated into the chromosomes and resulted in the regions termed genomic islands (GIs) (Koonpaew et al., 2000; Holden et al., 2004; Duangsonk et al., 2006). The repetitive genetic composition of *B. pseudomallei* genomes such as tandem repeats and transposons is believed to aid the integration of horizontally acquired elements into the chromosomes (Holden et al., 2004; U'Ren et al., 2007). Despite several plasmid elements being identified in *B. pseudomallei* genomes (Holden et al., 2004), four complete plasmids has been characterized so far (NCBI, December 2020) (Nandi et al., 2015). With long-read sequencing technology, the plasmid could be more reliably assembled, thereby allowing more genetic variations to be studied. At least 16 GIs had been identified in the K96243 genome (Holden et al., 2004). Several studies highlight distinct geographical distribution of GIs or their combinations (Ou et al., 2005; Duangsonk et al., 2006; Sim et al., 2008; Tuanyok et al., 2008), leading to a hypothesis that each GI or combination of GIs may confer different fitness under different environments (Sim et al., 2008). GIs carry several virulence factors, many of which are known to interact with hosts, including filamentous hemagglutinin adhesin (*fhaB3*), and a member of two-partner secretion system (*bpaAB*) (Sim et al., 2008).

GENETIC, STRUCTURAL OR MOLECULAR VARIATIONS OBSERVED IN HOST-PATHOGEN INTERACTION

Genetic, structural and molecular variations and their functional impact to *B. pseudomallei* - host interaction have been the center of melioidosis research over the past few decades. We will summarize genetic and structural diversity in bacterial genes known to mediate the infection (**Figure 2** and **Supplementary Table 1**). The genes discussed here are shown to be essential for infection by a genome-wide saturation mutagenesis (Moule et al., 2015) and also expressed during *in vivo* infection (Ooi et al., 2013). When host partners are known; variations detected in host proteins, and the host response are also described (**Figure 3**). It should be noted that there are large overlaps in the host immune system and the examples described here only represent a fraction of the whole machinery.

Genetic Variations in *B. pseudomallei* Flagella Systems and Host Toll-Like Receptor 5 (TLR5)

Flagella systems of *B. pseudomallei* enable the motility of the bacterium intracellularly as well as in the environment (**Figure 2**; DeShazer et al., 1997; French et al., 2011). Flagella filaments are constituted from thousands of flagellin protein monomers, encoded by *fliC* gene, strung into protofilaments before being braided to form a flagellum (Samatey et al., 2001). Flagellin has been known for its extreme diversity with 113,285 unique nucleotide sequences across the prokaryote phyla (Hu and Reeves, 2020). Flagellin inactivation in *B. thailandensis* resulted in the reduction of bacterium intracellular motility and cell-to-cell spread (French et al., 2011). Using PCR-restriction fragment length polymorphism analysis, four different alleles of flagellin protein (*fliC: BPSL3319*) were identified from 100 Malaysian clinical isolates (Tay et al., 2010), and more is expected if investigated with next-generation sequencing. Proteins in the flagella systems were reported to elicit host immune response, highlighting them as melioidosis vaccine candidates (Brett et al., 1994; Chua et al., 2003; Chuaygud et al., 2008; Gregory et al., 2015; Koosakulnirand et al., 2018).

In addition to flagella, a fimbrial gene cluster which is required for cell adherence also displays genetic variation with distinct alleles being predominantly detected in different geographical locations. A *yersinia*-like fimbrial (YLF) gene cluster, believed to be horizontally acquired was shown to be more prevalence in Southeast Asia and thus was used as a marker for the introduction of *B. pseudomallei* from this region (Tuanyok et al., 2007; Sarovich et al., 2014; Chewapreecha et al., 2017). A putative type I fimbrial protein *BPSL1626* was shown to elicit an immune response and have a potential as a vaccine candidate against melioidosis (Capelli et al., 2018). Whereas, *B. thailandensis*-like flagellum and chemotaxis (BTFC) gene cluster is believed to be an ancestral sequence in *B. pseudomallei* and is most common in Australia (Tuanyok et al., 2007). The YLF and BTFC gene clusters are mutually exclusive between the two endemic areas where the latter may implicate in the spread between cell-to-cell by the flagellar protein leading to the formation of multinucleated giant cells (MNGCs) and eventually apoptosis/cell death (French et al., 2011).

It has been established that flagellin is a ligand of TLR5 (**Figure 3**). The recognition stimulates pro-inflammatory responses including the rise in intracellular calcium ion and upregulation of pro-inflammatory cytokine TNF-α and IL-6 (Hayashi et al., 2001; Chen et al., 2007). TLR5 stop codon polymorphism, $TLR5_{1174C>T}$ or rs5744168 was strongly associated with protection against fatality as well as organ failure in a case-control cohort study. When challenged with *B. pseudomallei,* $TLR5_{1174C}$ could mediate the activation of NF-κB, while $TLR5_{1174T}$ could not; additionally, $TLR5_{1174T}$ saw reduced flagellin-induced cytokines levels (West et al., 2013). Another independent investigation on the same population also found the association between the truncated $TLR5_{1174C>T}$ variant and survival from acute melioidosis, as well as, a lower rate of bacteraemia (Chaichana et al., 2017). Furthermore, both studies also found $TLR5_{1174C>T}$ variant with a lower level of anti-inflammatory IL-10, which the authors suggested the possibility in which the mortality risk may be modulated by TLR5-driven IL-10 release. Interestingly, a recent study has suggested that the effect of this hypofunctional TLR5 variant may not be restricted to flagellin-driven pathway (Dickey et al., 2019). Another polymorphic variant is $TLR5_{1846T>C}$ which was also associated with protection against death and blunted flagellin-driven cytokine response; however, the authors also reported high linkage disequilibrium of the variant with TLR51174C>T (Chantratita et al., 2014), which might reduce the confidence of causal relationship.

Structural Variations in *B. pseudomallei* LPS and Genetic Variation in Host Toll-Like Receptor 2 and 4 (TLR2 and TLR4)

B. pseudomallei possesses an extensive network of polysaccharides on its outer membrane, namely capsular polysaccharide (CPS) and lipopolysaccharide (LPS) (**Figure 2**, for a full review on CPS and LPS in *Burkholderia* spp.; Cloutier et al., 2018). Both are known to play a vital role in virulence of melioidosis and have been used as subunits for vaccine development (Brett and Woods, 1996; Nelson et al., 2004; Wikraiphat et al., 2009; Scott et al., 2014b). Based on a cellular compartment, *B. pseudomallei* LPS can be divided into (Knirel et al., 1992): lipid A—an endotoxic component embedded in the phospholipid bilayer of the outer membrane; inner and outer core oligopolysaccharide; and O-antigen (Novem et al., 2009). Multiple *in vitro* studies showed LPS challenge could mount innate and adaptive immune responses, and Nitric Oxide (NO) production, with different LPS serotypes (A, B, and B2) reported to mount different magnitude of responses (Norris et al., 2017). Some genetic data of LPS biosynthesis operons were available. However, many of these were generated from short-read sequencing platforms, which are not ideal to investigate LPS due to the repetitive nature of the LPS locus. Long-read technologies, on the other hand, could

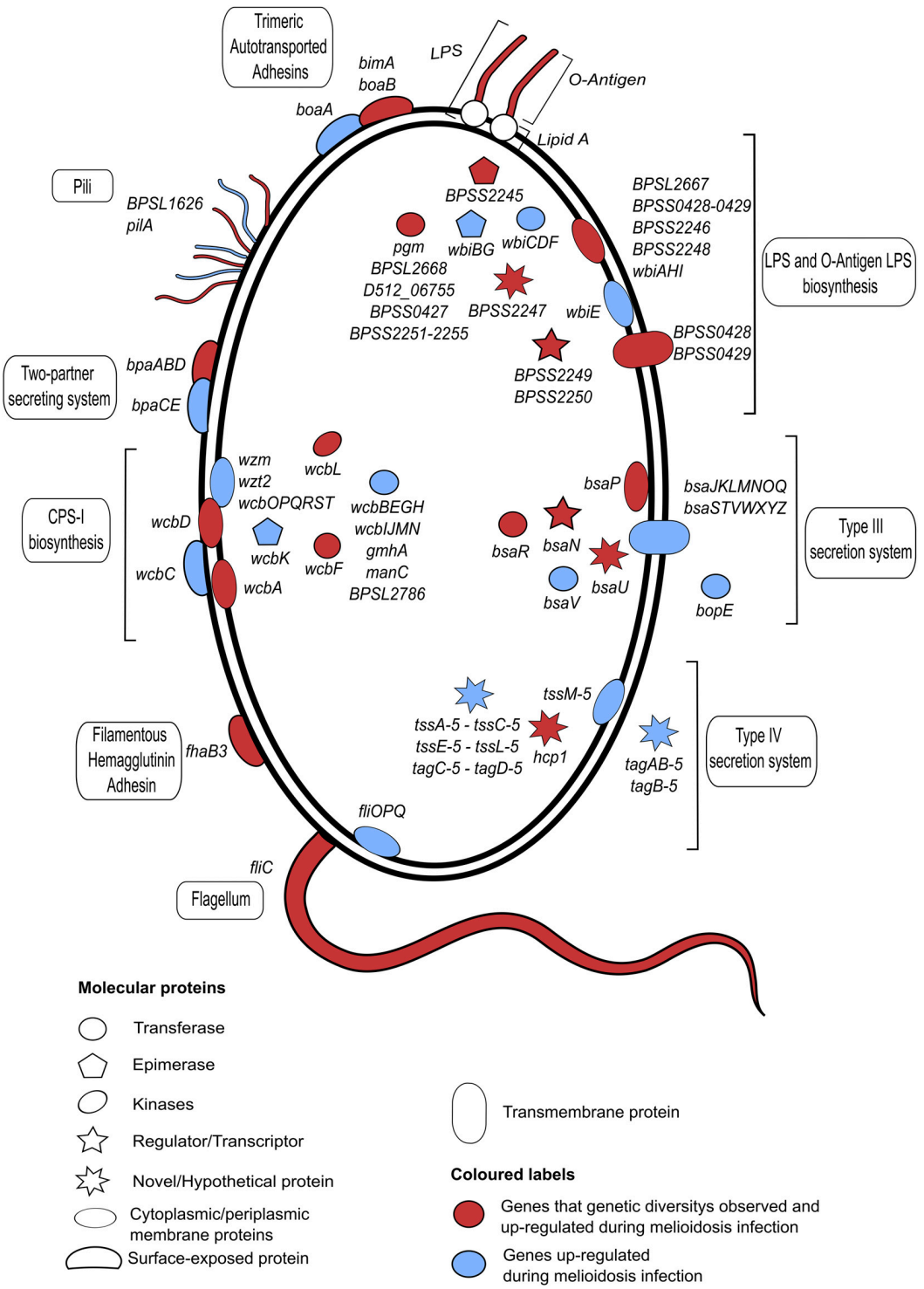


FIGURE 2 | A summary of *B. pseudomallei* genes or operons that are expressed during infection and characterized genetic variations. All plotted genes or operons are up-regulated during *in vivo* infections (Ooi et al., 2013), with a subset that reported genetic variations marked in red. The shape and location of each individual gene indicate the gene function and cellular compartment, respectively. All annotated genes or operons and their functions are described. The bacterium displays a repertoire of antigenic variations, including lipopolysaccharides (LPS), capsular polysaccharides (CPS) and surface proteins. *B. pseudomallei* LPS is immunologically classified into a number of serotypes A, B, and B2; with each serotype reported to be heterogeneously distributed across distinct geographical locations. Another highly diverse virulent protein is a fimbrial protein which displays a strong geographical distribution between Australia and Asia. Strains from Asia commonly possess a *Yersinia*-like fimbrial (YLF) gene cluster that believed to be horizontally acquired. *B. pseudomallei* carries 4 different types of CPS: CPS I, CPS II, CPS III, and CPS IV. A full gene description is provided in **Supplementary Table 1**.

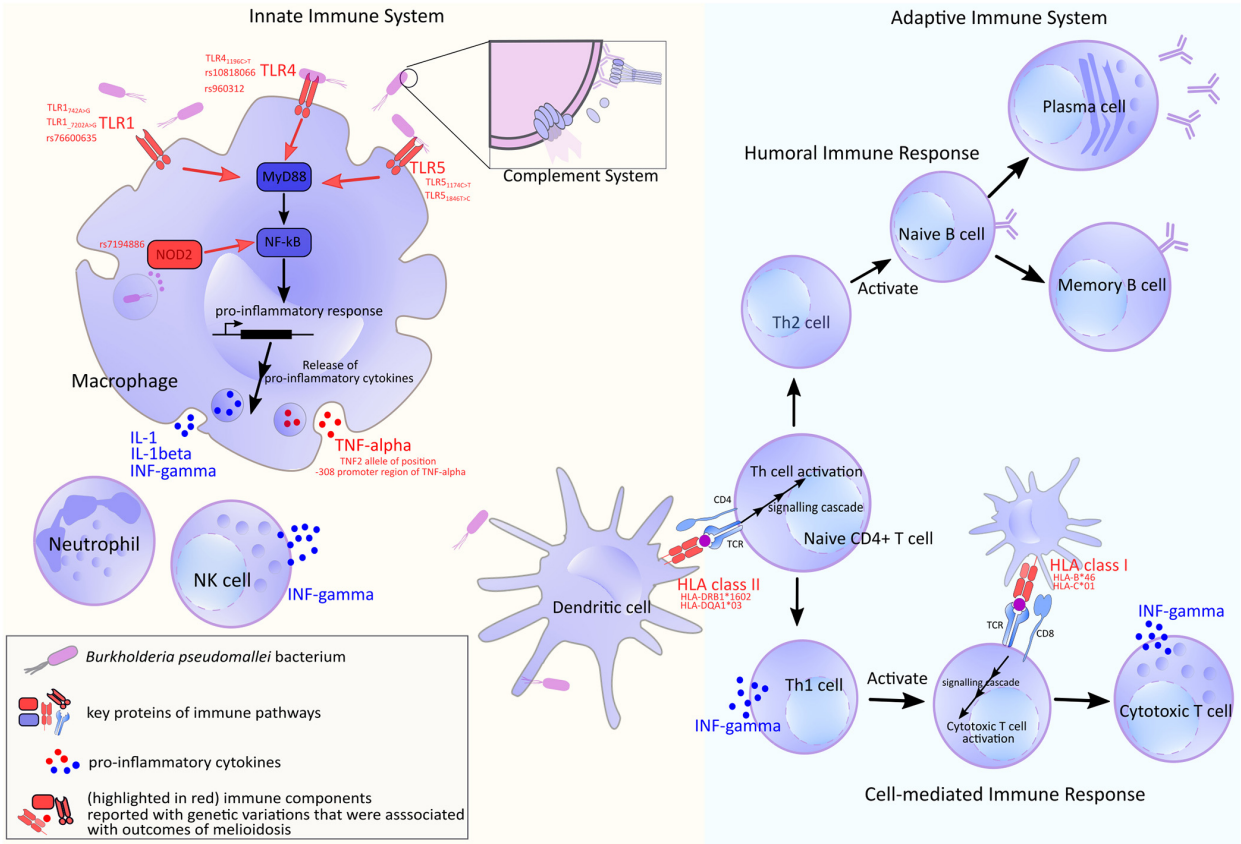


FIGURE 3 | A summary of human host immune components used in defending against *B. pseudomallei.* Genes or molecules that show genetic variations are marked in red. Upon infection, the bacterial antigens are recognized by host receptors including Toll-like receptor (TLR) and HLA which also display large genetic diversity. The latter is reported to be varied by ethnic groups (de Bakker et al., 2006; Gourraud et al., 2014). Following an entry into host, *B. pseudomallei* faces the innate axis of the immune system. Macrophages and neutrophils are recruited to the site early upon the infection; these cells are reported to be essential to the early bacterial containment and clearance (Breitbach et al., 2006; Easton et al., 2007), though excessive recruitment of neutrophils may have a negative outcome that allows *B. pseudomallei* to propagate intracellularly (Ceballos-Olvera et al., 2011). These innate phagocytes possess pattern recognition receptors (PRRs), such as surface receptors TLR2, TLR4, and TLR5, as well as, cytosolic receptors NOD2; these are reported to be vital for the fight against *B. pseudomallei* (West et al., 2008; Myers et al., 2014; Weehuizen et al., 2015; Birnie et al., 2019). The signals transduced by these receptors result in mobilization of nuclear factor NF-κB which trigger appropriate immune responses including the synthesis of pro-inflammatory cytokines and initiation of the downstream adaptive immune cascades (Pothlichet and Quintana-Murci, 2013). It is through the antigen-presenting cells and their appropriate antigen-HLA class II complexes that enable the activation of CD4 + T lymphocytes. The naïve CD4 + T cells sit at the central part of the adaptive axis. They can differentiate into Th1 cells which facilitate cell-mediated immune response by CD8 + cytotoxic lymphocytes. A study has shown that strong CD4 + and CD8 + T cell response was elicited during acute melioidosis and the lower cellular response was correlated to fatality (Jenjaroen et al., 2015). On the other hand, naïve CD4 + T cells can also mature to be Th2 cells which initiate class switching of B cells and support humoral immune response.

overcome this assembly issue, giving complete contigs of the bacterium genome.

To date, 4 serotypes of LPS were previously characterized. Type A LPS is a majority serotype found in both endemic areas, namely Southeast Asia and Australia (Anuntagool et al., 2000, 2006; Tuanyok et al., 2012). Whereas, serotype B is less common in the endemic areas (Australia and Southeast Asia) and is prevalent in both clinical and environmental origins (Tuanyok et al., 2012). However, a recent analysis of clinical *B. pseudomallei* isolates revealed that Serotype B was highly predominant in India (Shaw et al., 2019). LPS serotype B2 is a variant of LPS serotype B and more commonly found in Australia and Papua New Guinea (Tuanyok et al., 2012). In contrast, LPS R or Rough serotype lacks O-antigen moiety of the LPS structure. It was identified using SDS-PAGE and silver staining techniques with no O-antigen ladder pattern. Type R is relatively rare and found only from the Australian clinical and environmental strains (Anuntagool et al., 2006). Interestingly, it is frequently prevalent in the patients with a relapse history of melioidosis, however, there is no direct association between them (Limmathurotsakul et al., 2014b).

A comparative genomic analysis revealed that there are distinct variations in the core compositions of the O-antigen LPS biosynthesis gene clusters between A, B, and B2 serotypes. However, there are several genes conserved among them. In addition, type A LPS gene operon is also observed in both *B. thailandensis* and *B. mallei,* closely related *Burkholderia* species (Tuanyok et al., 2012). Interestingly, genomic analysis of clinical isolates from Madagascar revealed that there is a 13.5 kb deletion

observed in the LPS biosynthesis gene cluster of serotype B, conserving only some genes in the cluster that are essential for the biosynthesis of LPS B2 (Sarovich et al., 2016). The lack of some core genes in the LPS biosynthesis cluster could lead to the reduction of its serological properties. This is supported as some evidence suggested that the strains with LPS type B2 become sensitive to 30% normal human serum whereas the strains with LPS type B remain resistant (Tuanyok et al., 2012).

On the other hand, a single nucleotide insertion of the *wbiI* gene is observed in the LPS biosynthesis gene cluster of a rough serotype in a patient with more than 16-year chronic lung infection associated with melioidosis. This frame-shift mutation of *wbiI* gene (an essential gene for the O-antigen synthesis) disrupts the epimerase/dehydratase function of this gene and results in the loss of O-antigen moiety, possibly switching the serotype of *B. pseudomallei* isolates from type A to type R (Price et al., 2013; Pearson et al., 2020). Pearson and colleagues also revealed the nucleotide insertion of *D512_15771 (wbiH)* and *D512_06755* gene are observed in *B. pseudomallei* MSHR6686 of the same patient. The mutation in both genes may confer to the reduction in LPS modification and production, assisting in the escape from the host immune response. In addition, a partial deletion of *D512_20407, wbiA* homolog is also identified in MSHR6686 (Price et al., 2013; Pearson et al., 2020). However, genomic analysis of the rough serotype that is naturally found in the environment or initial infection has not been done to determine the genetic makeup of this serotype.

Not only genetic heterogeneity exists among the serotypes of LPS, further structural diversity is also observed at the O-antigen of LPS serotypes (Stone et al., 2012; Norris et al., 2017). The O-antigen is one of the LPS components with a structure of unbranched disaccharide repeat units. Remarkably, the structural modifications of these sugar chains were observed where substitutions of 2-O-methylated and 4-O-acetylated at talose residues were observed only in about 33% of the LPS serotype A. Whereas the rest bear 2-acetyl substituents at the same residues (Perry et al., 1995). More recently, a structural analysis of O-antigen in the serotype A reported that the modification of the talose residue is more complex than what was previously reported (Heiss et al., 2013). Although multiple gene inactivation studies revealed that *wbiA* and *oacA* genes are essential for these modifications at the talose residues (Brett et al., 2003, 2011), it is possible that structural diversities of the LPS O-antigen are further modified at post-transcriptional and post-translational levels, for instance, a length variation of O-antigen chain observed in *Escherichia coli* O9a which is tightly controlled by the biosynthetic enzymes dynamic (King et al., 2014).

The LPS is common across Gram-negative bacteria and also a well-established pathogen-associated molecular pattern (PAMP) that can trigger the pro-inflammatory innate immune response, such as translocation of NF-κB and TNF-α cytokine release, via its interaction with TLR4, CD-14 and adaptor protein MD-2 (Park and Lee, 2013). Although evidence has been inconclusive, LPS sensing was shown to be through TLR4 in murine models, and TLR4 as well as TLR2 (Wiersinga et al., 2007; West et al., 2008) in human models (**Figure 3**; Weehuizen et al., 2015). Polymorphisms of TLR4 were reported in humans with $TLR4_{1196C>T}$ allelic variant associated with protection against melioidosis when compared to non-hospitalized controls (West et al., 2012). In addition, TLR4 rs10828066 SNP variant was significantly associated to be protective against melioidosis in both adjusted non-hospitalized and hospitalized control groups, whereas TLR4 rs960312 was associated with bacteraemic or lung melioidosis (West et al., 2012).

Genetic Variations in *B. pseudomallei* CPS and Host B Cell Repertoire

CPS is one of the key virulence determinants in *B. pseudomallei* (**Figure 2**). To date, four different types of CPS (CPS I-IV) have been described (Holden et al., 2004). CPS-I biosynthesis cluster is a large 34.5 kb operon harboring on the chromosome one of *B. pseudomallei* (*wcbT* to *manC*). A study in an animal model reported that the CPS-I cluster is required for the full virulence of *B. pseudomallei* (Sarkar-Tyson et al., 2007). CPS-I exhibits an immunogenic role and the absence of this gene cluster resulted in the attenuation of *B. pseudomallei* in a mouse model (Atkins et al., 2002; Reckseidler-Zenteno et al., 2005; Parthasarathy et al., 2006). A passive protection was observed when murine intranasal infection models were immunized with anti-CPS monoclonal antibodies (AuCoin et al., 2012), with the later study reported that conjugated CPS could provide the highest degree of protection (Scott et al., 2014a).

The structural basis of CPS-I is highly conserved (Perry et al., 1995). Remarkably, CPS-I is believed to be horizontally transferred between species as the GC content of this CPS-I cluster is about 58% (Reckseidler et al., 2001). In addition, it is possible that the horizontal acquisition of this gene cluster was a key event in the pathogenic evolution of *B. pseudomallei* compared to *B. thailandensis* where the cluster is absent (Yu et al., 2006). Nevertheless, several genes involved in the biosynthesis of CPS-I of *B. pseudomallei* were reported in *B. thailandensis* with a relatively low sequence identity (75%). In addition, a novel variant of *B. thailandensis* strains (BTCV) is found to carry a *B. pseudomallei* like CPS-I (95% sequence identity) gene cluster with an identical organization. Although the origin of the CPS-I gene cluster in *B. pseudomallei* has yet been identified, several genes involved in the sugar biosynthesis appear to be orthologs to the genes identified in the Yersinia species *(Y. pseudotuberculosis H892/87)* and some in other gram-negative bacteria (Cuccui et al., 2012). It is possible that the CPSI gene cluster is not entirely acquired from one bacterial species but several organisms. Population studies also detected genetic variations in CPS-I (Challacombe et al., 2014; Chewapreecha et al., 2017). Whether this genetic variation has implications on virulence is a subject of further investigation. Another CPS-I variant could be found in *B. pseudomallei* isolated from a 16-year chronic melioidosis patient. A single nucleotide insertion is observed within *wcbR*, an important component involved in the fatty acid synthesis of CPS-I, causing a frameshift mutation in the earlier samples isolated from this patient (Price et al., 2013; Pearson et al., 2020). A CPS-I deletion region, which includes *wcbR* gene, has shown in the reduction of CPS production in hamster model but not absence entirely (Gutierrez and Warawa, 2016).

Price et al. (2013), suggested that this degree of CPS-I-dependent virulence decreases and may consequently be a critical step in the progression for melioidosis to become a chronic-carriage disease.

Unlike CPS-I, CPS-II and CPS-III have been associated with *B. pseudomallei* persistence in the environment (Reckseidler-Zenteno et al., 2009, 2010), while CPS-IV is less well-characterized. Up to date, no reports suggest the genetic variations of CPS-II, -III, and -IV biosynthesis clusters in *B. pseudomallei.* The diversity and function of these capsules in the pathogenesis, immunomodulation and environmental adaptation of *B. pseudomallei* warrant further studies.

We next considered host partners that interact with *B. pseudomallei* CPS. The CPS has been shown to elicit strong host immune response (Reckseidler et al., 2001; Atkins et al., 2002). CPSs from various *B. pseudomallei* strains were recognized by the same group of monoclonal antibodies which suggests a limited number of epitopes of this molecule (Zou et al., 2008). The molecule's monomers can cross-link to B cell receptors and, at an appropriate density, induce a downstream humoral response that is distinct from T cell-dependent pathway (Clarke et al., 2013; Akkaya et al., 2020). Antigen recognition through B-cell receptors is formed through random somatic changes of germline DNA. This results in a repertoire of distinct sequences that enable antigen recognition across wide ranges of pathogens. The B-cell receptors dynamic in melioidosis has not been studied.

Genetic Variations in *B. pseudomallei* Adhesins and Autotransporters

Filamentous Hemagglutinin Adhesin (FHA) of *B. pseudomallei* is highly diverse (**Figure 2**). Early investigation of genomic islands (GIs) from five *B. pseudomallei* strains identified three different *fhaB* gene clusters on different GIs of the bacteria. The bacterium could carry multiple *fhaB* gene clusters, many of with carrying either a combination of cluster I (GI5a/GI5a.1/GI5a.2) and cluster III (GI16/GI16.1), or cluster III alone (Tuanyok et al., 2008). More genetic variations of the FHA loci were reported from the Australian isolates (Chewapreecha et al., 2017). The *fhaB3* gene (*BPSS2053*) has been characterized as an important virulence factor of *B. pseudomallei*; enabling bacterial binding to the host epithelial cells (Sim et al., 2008), and an anti-macrophage factor (Dowling et al., 2010). It was found in all isolates from Thailand but found in only 83% of Australian strains where the absence of this gene in Australia population correlated with the skin abscess formation and lower mortality rate (Sarovich et al., 2014).

Autotransporter (AT) proteins are one of the largest family of the secretion systems in Gram-negative bacteria, allowing the transportation across the bacterial membrane as well as involving the virulence and immunogenicity of the pathogens (for review Lazar Adler et al., 2011). In *B. pseudomallei* K96243, a sequence analysis revealed to have at least 11 ATs located in the genome, including putative Trimeric Autotransporter Adhesins (TAAs), *bimA* (Stevens J.M. et al., 2005) and *boaB,* two-partner secretion system (TPS), *bpaA, bpaB*, and *bpaD* (Campos et al., 2013). Inactivation in some of these genes attenuated and reduced the intracellular survival of *B. pseudomallei* in the macrophage-like cells (Lazar Adler et al., 2015). ATs are variably present in the genomes. The gene *boaB (BPSL1705)*, which aids bacterial binding to host respiratory cells (Balder et al., 2010), was found to be absent in several strains in Africa, Brazil (Sarovich et al., 2016). When ATs are present, genetic variants can also be observed. Australian-specific and African-specific variants have been reported in *bpaA, bpaB,* and *bpaD (BPSS1434, BPSL2063,* and *BPSS0088,* respectively*)* (Brown et al., 2004; Tuanyok et al., 2008).

Several experiments demonstrated the role of BimA in actin polymerization and motility (Stevens M.P. et al., 2005; French et al., 2011), promoting a movement of cell-to-cell spread and within the host cells (Lazar Adler et al., 2011). Furthermore, BimA is antigenic (Suwannasaen et al., 2011) and seroactive (Felgner et al., 2009). Two variants of *bimA* (*BPSS1492*) were identified: *B. pseudomallei bimA* (*bimA*$_{BP}$), and *bimA*-like *B. mallei (bimA*$_{BM}$*)*. The latter is an ortholog of *bimA* from *Burkholderia mallei* (95% sequence identity) and displays the same domain organization: a single actin monomer binding motif (WH2), a proline-rich domain and a transmembrane anchor domain (Sitthidet et al., 2008). Whereas the *bimA*$_{BP}$ consists of two predicted WH2 domains, a proline-rich domain, a membrane anchor domain and an additional predicted casein kinase II (Stevens M.P. et al., 2005). Of note, the variation was also observed between the predicted proline-rich domain of the *bimA*$_{BM}$ and *B. mallei bimA* where the former has fewer motifs (Sitthidet et al., 2008). Distinct geographical distribution of *bimA*$_{BP}$ and *bimA*$_{BM}$ have been reported (Sitthidet et al., 2008; Sarovich et al., 2014; Shaw et al., 2019). Benanti et al. (2015) illustrated that both BimAs are functionally similar and associated with the nucleation and elongation of actin filaments with more plaque formation observed in *bimA*$_{BP}$ variant compared to *bimABM*. A subunit vaccine investigation reported a significant increased survival rate of BALB/c mice immunized with BimA$_{BM}$ recombinant, when the mice were challenged with both *B. mallei* and *B. pseudomallei* (Whitlock et al., 2010).

Genetic Variations in *B. pseudomallei* Secretion Systems

Several secretion systems and secreted proteins in *B. pseudomallei* are vital to combat the host immune defense and environmental stresses. *B. pseudomallei* has at least three type III secretion systems (T3SSs) (for review, Vander Broek and Stevens, 2017)and six type VI secretion systems (T6SSs).

Among T3SSs, T3SS-3 is better characterized and believed to be an integral part of the full *B. pseudomallei* virulence in mice and hamster models, as well as associated with the intracellular survival and dissemination. K-mer based sequence analysis has recently demonstrated that variations in the T3SS-3 gene cluster were detected across the global population of *B. pseudomallei* (Chewapreecha et al., 2017). This included *bsaU(BPSS1539), bsaR (BPSS1542), bsaP (BPSS1544), and bsaN* (BPSS1546) (Chewapreecha et al., 2017) which are believed to be involved in intracellular escape (Pilatz et al., 2006), predicted chaperone protein (Panina et al., 2005), T3SS-3 secretion regulator (Broek et al., 2015), and T3SS-3 regulator (Chen et al., 2014),

respectively. In addition, a partial deletion of genes in the T3SS is also observed in a patient with persisted infection of melioidosis as a result of within-host adaptations (Pearson et al., 2020). For T6SSs, T6SS-5 is better characterized (for review, Lennings et al., 2019)and functionally confirmed to mediate the translocation of effector proteins via contact-dependent manner (Silverman et al., 2012) and MNGC formation (Burtnick et al., 2011; Chen et al., 2011). In addition, an experimental study suggested that T6SS-5 may involve in the intracellular survival of *B. pseudomallei* in macrophages where the expression level of T6SS-5 is dependent on *virAG and bprC* regulatory gene (Chen et al., 2011). Genetic variations are also observed in several genes clustered T6SS-5 with distinct geographical distribution (Chewapreecha et al., 2017). Remarkably, a large deletion region on chromosome 2 of an environmental *B. pseudomallei* A4 and an isogenic strain of K96243 has been demonstrated of which they failed to form a plaque in epithelium cells. Of those, the deleted genes are including genes involved in T3SS-3 and T6SS-5 systems. Interestingly, Saiprom et al. (2020) also identified an absence of multiple T3SS-1 genes of previously described Thai environmental strain RF80.

Genetic Variations in Human Leukocyte Antigen (HLA)

For the rest of the bacterial factors that do not get recognized by any specific innate pattern recognition receptors, these antigens will be processed and presented on HLA of antigen-presenting cells such as dendritic cells before getting recognized by appropriate T cell receptors (TCR) of the adaptive immune axis; this HLA-peptide-TCR interaction kicks start the adaptive immune response. There are three different classes of HLA: class I interacts with TCRs on CD8 + T cells while class II binds to TCRs on CD4 + T cells, and the less well-established class III which is not involved in antigen processing and presentation (Dendrou et al., 2018). Acute melioidosis patients with diabetes mellitus were reported to have lower HLA-DR expression on plasmacytoid dendritic cells than the non-diabetic diseased group (Kronsteiner et al., 2019). In addition, in non-diabetic patients, fatal cases presented with significantly lower expression of HLA-DR on monocytes and plasmacytoid dendritic cells, compared to the survived cases (Kronsteiner et al., 2019). Similar observation was found in murine infection models where *B. pseudomallei*-infected BALB/c and C57BL/6 mice showed reduced expression of MHC class II on plasmacytoid dendritic cells, however, this was not statistically significant (Williams et al., 2015).

Genetic variations in HLA have been linked to the melioidosis outcomes. A work conducted in an endemic area of Thailand has compared HLA allele frequencies in melioidosis cases and healthy controls. The authors reported a significant increase of DRB1*1602 frequency in melioidosis patients, compared to healthy controls. Moreover, an increase in HLA-DRB1*1602 and a decrease in HLA-DQA1*03 allele frequencies were associated with septicaemic cases of melioidosis (Dharakul et al., 1998). Another study conducted on the same Thai population screened a panel of various HLA class I genotype frequencies in survived and fatal cases of melioidosis. The authors found that HLA-B*46 and HLA-C*01 were associated with increased mortality; they were also reported to be in linkage disequilibrium (Dunachie et al., 2017). Interestingly, HLA has been linked with diabetes mellitus, the prominent comorbidity of melioidosis with 12-fold increased risk of developing the disease (Limmathurotsakul et al., 2010a). Several recent publications using GWAS have identified variations of HLA and their corresponding protective or predisposing association with type 2 diabetes (Williams et al., 2011; Scott et al., 2017; Zhao et al., 2017). These prompt further characterization of HLA variants and their relationship with melioidosis progression when modulated by patient diabetic status. However, genetic studies on HLA and melioidosis have suffered from small sample size and availability of reliable HLA typing platform. This is largely impeded by the under-representation of genomic data from the population from melioidosis endemic areas. The availability of data is crucial as this improves imputations and discovery of new causal variants and disease association.

DISCUSSION

The advancement in omic technologies has improved our understanding of co-evolution between *B. pseudomallei* and different hosts, thereby guiding better control policy, treatment option and vaccine design. Proteins or molecules that participate in host—*B. pseudomallei* interaction are extremely variable. Their variations have been seen at the genomic, epigenetic, transcriptomic and proteomic levels. Although genome data for human population from melioidois endemic areas is still scarce, genome data for *B.pseudomallei* has been accumulating (Holden et al., 2004; Hayden et al., 2012; Price et al., 2013, 2016; Sahl et al., 2013, 2016; Daligault et al., 2014; Bugrysheva et al., 2015; Chen et al., 2015; Hsueh et al., 2015; Johnson et al., 2015a,b; McRobb et al., 2015; Nandi et al., 2015; Sidjabat et al., 2015; Song et al., 2015; Spring-Pearson et al., 2015; Viberg et al., 2015; Chapple et al., 2016; Aziz et al., 2017; Chewapreecha et al., 2017, 2019; Podnecky et al., 2017; Viberg et al., 2017; Webb et al., 2019). When combined with spatial and temporal information, this allows further exploration of allelic variants and a shift in allele frequency over space and time. Moreover, a genome-wide saturation mutagenesis can aid prediction of essential genes required under certain conditions (Moule et al., 2015). This can be coupled with transcriptome information to understand variations in the expression patterns (Chieng et al., 2012; Ooi et al., 2013; Price et al., 2018) during the course of infection, and across different host types. A dual host-pathogen transcriptome study has not been conducted for melioidosis but is promising to provide valuable insight into the interaction as well as variations that lead to different infection outcomes.

In this article, we mainly explored variations of *B. pseudomallei* genes implicated in human host infections at the genetic level. Many of these genes display a strong geographical signal which could either be a result of a founder effect following a migration out of Australia, or an acquisition of new alleles required for local adaptation after an introduction to new geographical location. For each virulence gene, we

also noted co-existence of multiple alleles in *B. pseudomallei* population isolated from the same geographical region. The genetic polymorphism could be maintained by balancing selection where each co-existing allele must be favored under different condition. In the context of virulent genes, these could involve competitions with different soil organisms. Signals for positive selection could be detected in genes that promote bacterial survival under hostile environment, and genes required for cell entry and adaptation to an intracellular lifestyle. The latter can be grouped as virulence genes although they may primarily be used in amoeba hosts rather than mammalian hosts. For multi-host—pathogen interaction, it is thus essential to consider virulence in a broad host and environmental context. The picture is far from complete at the moment, but more incoming omic data is promising to shed light on this complex relationship.

AUTHOR CONTRIBUTIONS

CCe conceived the study and performed data analyses. CCo and PB performed literature review. AS performed data analyses. All authors wrote the manuscript.

FUNDING

CCo was funded by Wellcome International Master Fellowship (221418/Z/20/Z). CCe was funded by Wellcome International Intermediate Fellowship (216457/Z/19/Z) and Sanger International Fellowship and Thailand Research Fund (MRG6280226).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.612568/full#supplementary-material

Supplementary Text | We estimated the proportion of infection outcome variations (survival vs. mortal outcomes after 28 days) explained by *B. pseudomallei* nucleotide polymorphisms (SNPs). SNPs were called from Chewapreecha et al. (2019, 2020) using kSNP v3 software. A filtering for minimum allele frequency at 0.01 (MAF = 0.01) resulted in 65,379 SNPs for the analysis. *B. pseudomallei* population was defined by PopPUNK (Lees et al., 2019). We employed the Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011) to calculate h^2 using genomic restricted maximum likelihood (GREML) while adjusting for population structure using population clusters defined by PopPUNK.

REFERENCES

Abbink, F. C., Orendi, J. M., and de Beaufort, A. J. (2001). Mother-to-child transmission of *Burkholderia pseudomallei*. *N. Engl. J. Med.* 344, 1171–1172.

Adams, N. M., Grassmann, S., and Sun, J. C. (2020). Clonal expansion of innate and adaptive lymphocytes. *Nat. Rev. Immunol.* 20, 694–707. doi: 10.1038/s41577-020-0307-4

Akkaya, M., Kwak, K., and Pierce, S. K. (2020). B cell memory: builing two walls of protection against pathogens. *Nat. Rev. Immunol.* 20, 229–238. doi: 10.1038/s41577-019-0244-2

Anuntagool, N., Aramsri, P., Panichakul, T., Wuthiekanun, V. R., Kinoshita, R., White, N. J., et al. (2000). Antigenic heterogeneity of lipopolysaccharide among *Burkholderia pseudomallei* clinical isolates. *Southeast Asian J. Trop. Med. Public Health* 31(Suppl. 1), 146–152.

Anuntagool, N., Wuthiekanun, V., White, N. J., Currie, B. J., Sermswan, R. W., Wongratanacheewin, S., et al. (2006). Lipopolysaccharide heterogeneity among *Burkholderia pseudomallei* from different geographic and clinical origins. *Am. J. Trop. Med. Hyg.* 74, 348–352. doi: 10.4269/ajtmh.2006.74.348

Atkins, T., Prior, R., Mack, K., Russell, P., Nelson, M., Prior, J., et al. (2002). Characterisation of an acapsular mutant of *Burkholderia pseudomallei* identified by signature tagged mutagenesis. *J. Med. Microbiol.* 51, 539–553. doi: 10.1099/0022-1317-51-7-539

AuCoin, D. P., Reed, D. E., Marlenee, N. L., Bowen, R. A., Thorkildson, P., Judy, B. M., et al. (2012). Polysaccharide specific monoclonal antibodies provide passive protection against intranasal challenge with *Burkholderia pseudomallei*. *PLoS One* 7:e35386. doi: 10.1371/journal.pone.0035386

Aziz, A., Currie, B. J., Mayo, M., Sarovich, D. S., and Price, E. P. (2020). Comparative genomics confirms a rare melioidosis human-to-human transmission event and reveals incorrect phylogenomic reconstruction due to polyclonality. *Microb. Genom.* 6:e000326.

Aziz, A., Sarovich, D. S., Harris, T. M., Kaestli, M., McRobb, E., Mayo, M., et al. (2017). Suspected cases of intracontinental *Burkholderia pseudomallei* sequence type homoplasy resolved using whole-genome sequencing. *Microb. Genom.* 3:e000139. doi: 10.1099/mgen.0.000139

Balder, R., Lipski, S., Lazarus, J. J., Grose, W., Wooten, R. M., Hogan, R. J., et al. (2010). Identification of *Burkholderia mallei* and *Burkholderia pseudomallei* adhesins for human respiratory epithelial cells. *BMC Microbiol.* 10:250. doi: 10.1186/1471-2180-10-250

Benanti, E. L., Nguyen, C. M., and Welch, M. D. (2015). Virulent *Burkholderia* species mimic host actin polymerases to drive actin-based motility. *Cell* 161, 348–360. doi: 10.1016/j.cell.2015.02.044

Birnie, E., Weehuizen, T. A. F., Lankelma, J. M., de Jong, H. K., Koh, G. C. K. W., van Lieshout, M. H. P., et al. (2019). Role of toll-like receptor 5 (TLR5) in experimental melioidosis. *Infect. Immun.* 87, e409–e418.

Bot, A., and Benites, J. (2005). *The Importance of Soil Organic Matter: key to Drought-Resistant Soil and Sustained Food Production*. Rome: Food and Agriculture Organization of the United Nations, 78.

Breitbach, K., Klocke, S., Tschernig, T., van Rooijen, N., Baumann, U., and Steinmetz, I. (2006). Role of inducible nitric oxide synthase and NADPH oxidase in early control of *Burkholderia pseudomallei* infection in mice. *Infect. Immun.* 74, 6300–6309. doi: 10.1128/iai.00966-06

Brett, P. J., Burtnick, M. N., Heiss, C., Azadi, P., DeShazer, D., Woods, D. E., et al. (2011). *Burkholderia thailandensis* oacA mutants facilitate the expression of *Burkholderia mallei*-like O polysaccharides. *Infect. Immun.* 79, 961–969. doi: 10.1128/iai.01023-10

Brett, P. J., Burtnick, M. N., and Woods, D. E. (2003). The wbiA locus is required for the 2-O-acetylation of lipopolysaccharides expressed by *Burkholderia pseudomallei* and *Burkholderia thailandensis*. *FEMS Microbiol. Lett.* 218, 323–328.

Brett, P. J., Mah, D. C., and Woods, D. E. (1994). Isolation and characterization of *Pseudomonas pseudomallei* flagellin proteins. *Infect Immun.* 62, 1914–1919. doi: 10.1128/iai.62.5.1914-1919.1994

Brett, P. J., and Woods, D. E. (1996). Structural and immunological characterization of *Burkholderia pseudomallei* O-polysaccharide-flagellin protein conjugates. *Infect. Immun.* 64, 2824–2828. doi: 10.1128/iai.64.7.2824-2828.1996

Brock, D. A., Read, S., Bozhchenko, A., Queller, D. C., and Strassmann, J. E. (2013). Social amoeba farmers carry defensive symbionts to protect and privatize their crops. *Nat. Commun.* 4:2385.

Broek, C. W. V., Chalmers, K. J., Stevens, M. P., and Stevens, J. M. (2015). Quantitative proteomic analysis of *Burkholderia pseudomallei* Bsa type III secretion system effectors using hypersecreting mutants. *Mol. Cell. Proteom.* 14, 905–916. doi: 10.1074/mcp.m114.044875

Brown, N. F., Logue, C. A., Boddey, J. A., Scott, R., Hirst, R. G., and Beacham, I. R. (2004). Identification of a novel two-partner secretion system from *Burkholderia pseudomallei*. *Mol. Genet. Genom.* 272, 204–215. doi: 10.1007/s00438-004-1039-z

Bugrysheva, J. V., Sue, D., Hakovirta, J., Loparev, V. N., Knipe, K., Sammons, S. A., et al. (2015). Finished Annotated Genome Sequence of *Burkholderia pseudomallei* Strain Bp1651, a Multidrug-Resistant Clinical Isolate. *Genome Announc.* 3:e01427-15.

Burtnick, M. N., Brett, P. J., Harding, S. V., Ngugi, S. A., Ribot, W. J., Chantratita, N., et al. (2011). The Cluster 1 type VI secretion system is a major virulence determinant in *Burkholderia pseudomallei*. *Infect. Immun.* 79, 1512–1525. doi: 10.1128/iai.01218-10

Campos, C. G., Byrd, M. S., and Cotter, P. A. (2013). Functional characterization of *Burkholderia pseudomallei* trimeric autotransporters. *Infect. Immun.* 81, 2788–2799. doi: 10.1128/iai.00526-13

Capelli, R., Peri, C., Villa, R., Nithichanon, A., Conchillo-Solé, O., Yero, D., et al. (2018). BPSL1626: reverse and structural vaccinology reveal a novel candidate for vaccine design against *Burkholderia pseudomallei*. *Antibodies (Basel)* 7:26. doi: 10.3390/antib7030026

Ceballos-Olvera, I., Sahoo, M., Miller, M. A., Del Barrio, L., and Re, F. (2011). Inflammasome-dependent pyroptosis and IL-18 protect against *Burkholderia pseudomallei* lung infection while IL-1β is deleterious. *PLoS Pathog.* 7:e1002452. doi: 10.1371/journal.ppat.1002452

Chaichana, P., Chantratita, N., Brod, F., Koosakulnirand, S., Jenjaroen, K., Chumseng, S., et al. (2017). A nonsense mutation in TLR5 is associated with survival and reduced IL-10 and TNF-alpha levels in human melioidosis. *PLoS Negl. Trop. Dis.* 11:e0005587. doi: 10.1371/journal.pntd.0005587

Challacombe, J. F., Stubben, C. J., Klimko, C. P., Welkos, S. L., Kern, S. J., Bozue, J. A., et al. (2014). Interrogation of the *Burkholderia pseudomallei* genome to address differential virulence among isolates. *PLoS One* 9:e115951. doi: 10.1371/journal.pone.0115951

Chantratita, N., Tandhavanant, S., Myers, N. D., Chierakul, W., Robertson, J. D., Mahavanakul, W., et al. (2014). Screen of whole blood responses to flagellin identifies TLR5 variation associated with outcome in melioidosis. *Genes Immun.* 15, 63–71. doi: 10.1038/gene.2013.60

Chapple, S. N. J., Sarovich, D. S., Holden, M. T. G., Peacock, S. J., Buller, N., Golledge, C., et al. (2016). Whole-genome sequencing of a quarter-century melioidosis outbreak in temperate Australia uncovers a region of low-prevalence endemicity. *Microb. Genom.* 2:e000067.

Chen, Y., Schröder, I., French, C. T., Jaroszewicz, A., Yee, X. J., Teh, B.-E., et al. (2014). Characterization and analysis of the *Burkholderia pseudomallei* BsaN virulence regulon. *BMC Microbiol.* 14:206. doi: 10.1186/s12866-014-0206-6

Chen, Y., Wong, J., Sun, G. W., Liu, Y., Tan, G.-Y. G., and Gan, Y.-H. (2011). Regulation of type VI secretion system during *Burkholderia pseudomallei* infection. *Infect. Immun.* 79, 3064–3073. doi: 10.1128/iai.05148-11

Chen, Y. L., Chen, Y. S., Lin, H. H., Chan, C. W., Chen, S. C., and Chen, C. H. (2007). Immunostimulatory flagellin from *Burkholderia pseudomallei* effects on an increase in the intracellular calcium concentration and up-regulation of TNF-alpha by mononuclear cells. *Microbiol. Immunol.* 51, 81–86. doi: 10.1111/j.1348-0421.2007.tb03893.x

Chen, Y.-S., Lin, H.-H., Hsueh, P.-T., Liu, P.-J., Ni, W.-F., Chung, W.-C., et al. (2015). Whole-genome sequence of an epidemic strain of *Burkholderia pseudomallei* vgh07 in Taiwan. *Genome Announc.* 3:e00345-15.

Chewapreecha, C., Harris, S. R., Croucher, N. J., Turner, C., Marttinen, P., Cheng, L., et al. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* 46, 305–309. doi: 10.1038/ng.2895

Chewapreecha, C., Holden, M. T. G., Vehkala, M., Välimäki, N., Yang, Z., Harris, S. R., et al. (2017). Global and regional dissemination and evolution of *Burkholderia pseudomallei*. *Nat. Microbiol.* 2, 1–8.

Chewapreecha, C., Mather, A. E., Harris, S. R., Hunt, M., Holden, M. T. G., Chaichana, C., et al. (2019). Genetic variation associated with infection and the environment in the accidental pathogen *Burkholderia pseudomallei*. *Commun. Biol.* 2, 1–11.

Chewapreecha, C., Pensar, J., Chattagul, S., Pesonen, M., Sangphukieo, A., Boonklang, P., et al. (2020). Co-evolutionary signals from *Burkholderia pseudomallei* population genomics highlight its survival strategy in a hostile environment. *bioRxiv*[Preprint] doi: 10.1101/2020.08.11.245894

Chieng, S., Carreto, L., and Nathan, S. (2012). *Burkholderia pseudomallei* transcriptional adaptation in macrophages. *BMC Genom.* 13:328. doi: 10.1186/1471-2164-13-328

Chua, K. L., Chan, Y. Y., and Gan, Y. H. (2003). Flagella are virulence determinants of *Burkholderia pseudomallei*. *Infect. Immun.* 71, 1622–1629. doi: 10.1128/iai.71.4.1622-1629.2003

Chuaygud, T., Tungpradabkul, S., Sirisinha, S., Chua, K. L., and Utaisincharoen, P. (2008). A role of *Burkholderia pseudomallei* flagella as a virulent factor. *Trans. R. Soc. Trop. Med. Hyg.* 102(Suppl. 1), S140–S144.

Clarke, E. T., Williams, N. A., Findlow, J., Borrow, R., Heyderman, R. S., and Finn, A. (2013). Polysaccharide-specific memory B cells generated by conjugate vaccines in humans conform to the CD27+IgG+ isotype-switched memory B Cell phenotype and require contact-dependent signals from bystander T cells activated by bacterial proteins to differentiate into plasma cells. *J. Immunol.* 191, 6071–6083. doi: 10.4049/jimmunol.1203254

Cloutier, M., Muru, K., Ravicoularamin, G., and Gauthier, C. (2018). Polysaccharides from *Burkholderia* species as targets for vaccine development, immunomodulation and chemical synthesis. *Nat. Prod. Rep.* 35, 1251–1293. doi: 10.1039/c8np00046h

Cooke, G. S., and Hill, A. V. (2001). Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* 2, 967–977.

Cuccui, J., Milne, T. S., Harmer, N., George, A. J., Harding, S. V., Dean, R. E., et al. (2012). Characterization of the *Burkholderia pseudomallei* K96243 capsular polysaccharide I coding region. *Infect. Immun.* 80, 1209–1221. doi: 10.1128/iai.05805-11

Currie, B. J., and Jacups, S. P. (2003). Intensity of rainfall and severity of melioidosis, Australia. *Emerg. Infect. Dis. J. CDC* 9, 1538–1542. doi: 10.3201/eid0912.020750

Daligault, H. E., Davenport, K. W., Minogue, T. D., Bishop-Lilly, K. A., Broomall, S. M., Bruce, D. C., et al. (2014). Whole-genome assemblies of 56 *Burkholderia* species. *Genome Announc.* 2:e01106-14.

David, S., Sanchez-Buso, L., Harris, S. R., Marttinen, P., Rusniok, C., Buchrieser, C., et al. (2017). Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*. *PLoS Genet.* 13:e1006855. doi: 10.1371/journal.pgen.1006855

de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* 38, 1166–1172. doi: 10.1038/ng1885

Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. (2018). HLA variation and disease. *Nat. Rev. Immunol.* 18, 325–339.

DeShazer, D., Brett, P. J., Carlyon, R., and Woods, D. E. (1997). Mutagenesis *of Burkholderia pseudomallei* with Tn5-OT182: isolation of motility mutants and molecular characterization of the flagellin structural gene. *J. Bacteriol.* 179, 2116–2125. doi: 10.1128/jb.179.7.2116-2125.1997

Dharakul, T., Vejbaesya, S., Chaowagul, W., Luangtrakool, P., Stephens, H. A., and Songsivilai, S. (1998). HLA-DR and -DQ associations with melioidosis. *Hum. Immunol.* 59, 580–586. doi: 10.1016/s0198-8859(98)00052-4

Dickey, A. K., Chantratita, N., Tandhavanant, S., Ducken, D., Lovelace-Macon, L., Seal, S., et al. (2019). Flagellin-independent effects of a Toll-like receptor 5 polymorphism in the inflammatory response to *Burkholderia pseudomallei*. *PLoS Neg. Trop. Dis.* 13:e0007354. doi: 10.1371/journal.pntd.0007354

DiSalvo, S., Haselkorn, T. S., Bashir, U., Jimenez, D., Brock, D. A., Queller, D. C., et al. (2015). *Burkholderia* bacteria infectiously induce the proto-farming symbiosis of Dictyostelium amoebae and food bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5029–E5037.

Dowling, A. J., Wilkinson, P. A., Holden, M. T. G., Quail, M. A., Bentley, S. D., Reger, J., et al. (2010). Genome-wide analysis reveals loci encoding anti-macrophage factors in the human pathogen *Burkholderia pseudomallei* K96243. *PLoS One* 5:e15693. doi: 10.1371/journal.pone.0015693

Duangsonk, K., Gal, D., Mayo, M., Hart, C. A., Currie, B. J., and Winstanley, C. (2006). Use of a variable amplicon typing scheme reveals considerable variation in the accessory genomes of isolates of *Burkholderia pseudomallei*. *J. Clin. Microbiol.* 44, 1323–1334. doi: 10.1128/jcm.44.4.1323-1334.2006

Dunachie, S. J., Jenjaroen, K., Reynolds, C. J., Quigley, K. J., Sergeant, R., Sumonwiriya, M., et al. (2017). Infection with *Burkholderia pseudomallei* – immune correlates of survival in acute melioidosis. *Sci. Rep.* 7:12143.

Easton, A., Haque, A., Chu, K., Lukaszewski, R., and Bancroft, G. J. A. (2007). Critical role for neutrophils in resistance to experimental infection with *Burkholderia pseudomallei*. *J. Infect. Dis.* 195, 99–107. doi: 10.1086/509810

Ewbank, D. (2016). Measuring selection in human populations using the growth rate per generation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20150148. doi: 10.1098/rstb.2015.0148

Felgner, P. L., Kayala, M. A., Vigil, A., Burk, C., Nakajima-Sasaki, R., Pablo, J., et al. (2009). A *Burkholderia pseudomallei* protein microarray reveals serodiagnostic and cross-reactive antigens. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13499–13504. doi: 10.1073/pnas.0812080106

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235. doi: 10.1038/ng.3404

French, C. T., Toesca, I. J., Wu, T.-H., Teslaa, T., Beaty, S. M., Wong, W., et al. (2011). Dissection of the *Burkholderia* intracellular life cycle using a photothermal nanoblade. *Proc.Natl. Acad. Sci. U.S.A.* 108, 12095–12100. doi: 10.1073/pnas.1107183108

Gourraud, P.-A., Khankhanian, P., Cereb, N., Yang, S. Y., Feolo, M., Maiers, M., et al. (2014). HLA diversity in the 1000 genomes dataset. *PLoS One* 9:e97282. doi: 10.1371/journal.pone.0097282

Gregory, A. E., Judy, B. M., Qazi, O., Blumentritt, C. A., Brown, K. A., Shaw, A. M., et al. (2015). A gold nanoparticle-linked glycoconjugate vaccine against *Burkholderia mallei*. *Nanomedicine* 11, 447–456. doi: 10.1016/j.nano.2014.08.005

Gutierrez, M. G., and Warawa, J. M. (2016). Attenuation of a select agent-excluded *Burkholderia pseudomallei* capsule mutant in hamsters. *Acta Trop.* 157, 68–72. doi: 10.1016/j.actatropica.2015.12.006

Hantrakun, V., Rongkard, P., Oyuchua, M., Amornchai, P., Lim, C., Wuthiekanun, V., et al. (2016). Soil nutrient depletion is associated with the presence of *Burkholderia pseudomallei*. *Appl. Environ. Microbiol.* 82, 7086–7092.

Haselkorn, T. S., DiSalvo, S., Miller, J. W., Bashir, U., Brock, D. A., Queller, D. C., et al. (2019). The specificity of *Burkholderia* symbionts in the social amoeba farming symbiosis: prevalence, species, genetic and phenotypic diversity. *Mol. Ecol.* 28, 847–862. doi: 10.1111/mec.14982

Hayashi, F., Smith, K. D., Ozinsky, A., Hawn, T. R., Yi, E. C., Goodlett, D. R., et al. (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* 410, 1099–1103. doi: 10.1038/35074106

Hayden, H. S., Lim, R., Brittnacher, M. J., Sims, E. H., Ramage, E. R., Fong, C., et al. (2012). Evolution of *Burkholderia pseudomallei* in recurrent melioidosis. *PLoS One*.7:e36507. doi: 10.1371/journal.pone.0036507

Heiss, C., Burtnick, M. N., Roberts, R. A., Black, I., Azadi, P., and Brett, P. J. (2013). Revised structures for the predominant O-polysaccharides expressed by *Burkholderia pseudomallei* and *Burkholderia mallei*. *Carbohydr. Res.* 381, 6–11. doi: 10.1016/j.carres.2013.08.013

Holden, M. T. G., Titball, R. W., Peacock, S. J., Cerdeño-Tárraga, A. M., Atkins, T., Crossman, L. C., et al. (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14240–14245.

Hou, K., Burch, K. S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., et al. (2019). Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* 51, 1244–1251. doi: 10.1038/s41588-019-0465-0

Hsueh, P.-T., Chen, Y.-S., Lin, H.-H., Liu, P.-J., Ni, W.-F., Liu, M.-C., et al. (2015). Comparison of whole-genome sequences from two colony morphovars of *Burkholderia pseudomallei*. *Genome Announc.* 3:e01194-15.

Hu, D., and Reeves, P. R. (2020). The remarkable dual-level diversity of prokaryotic flagellins. *mSystems* 5:e00705-19.

Inglis, T. J., Rigby, P., Robertson, T. A., Dutton, N. S., Henderson, M., and Chang, B. J. (2000). Interaction between *Burkholderia pseudomallei* and *Acanthamoeba* species results in coiling phagocytosis, endamebic bacterial survival, and escape. *Infect. Immun.* 68, 1681–1686. doi: 10.1128/iai.68.3.1681-1686.2000

Inglis, T. J., Robertson, T., Woods, D. E., Dutton, N., and Chang, B. J. (2003). Flagellum-mediated adhesion by *Burkholderia pseudomallei* precedes invasion of *Acanthamoeba astronyxis*. *Infect. Immun.* 71, 2280–2282. doi: 10.1128/iai.71.4.2280-2282.2003

Jenjaroen, K., Chumseng, S., Sumonwiriya, M., Ariyaprasert, P., Chantratita, N., Sunyakumthorn, P., et al. (2015). T-Cell responses are associated with survival in acute melioidosis patients. *PLoS Neg. Trop. Dis.* 9:e0004152. doi: 10.1371/journal.pntd.0004152

Jia, G., Li, Y., Zhang, H., Chattopadhyay, I., Boeck Jensen, A., Blair, D. R., et al. (2019). Estimating heritability and genetic correlations from large health datasets in the absence of genetic data. *Nat. Commun.* 10:5508.

Johnson, S. L., Baker, A. L., Chain, P. S., Currie, B. J., Daligault, H. E., Davenport, K. W., et al. (2015a). Whole-Genome Sequences of 80 Environmental and Clinical Isolates of *Burkholderia pseudomallei*. *Genome Announc.* 3:e01282-14.

Johnson, S. L., Bishop-Lilly, K. A., Ladner, J. T., Daligault, H. E., Davenport, K. W., Jaissle, J., et al. (2015b). Complete genome sequences for 59 *burkholderia* isolates, both pathogenic and near neighbor. *Genome Announc.* 3:e00159-15.

Kariuki, S. N., Marin-Menendez, A., Introini, V., Ravenhill, B. J., Lin, Y. C., Macharia, A., et al. (2020). Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature.* 585, 579–583. doi: 10.1038/s41586-020-2726-6

Karlsson, E. K., Kwiatkowski, D. P., and Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* 15, 379–393. doi: 10.1038/nrg3734

Khojandi, N., Haselkorn, T. S., Eschbach, M. N., Naser, R. A., and DiSalvo, S. (2019). Intracellular *Burkholderia* Symbionts induce extracellular secondary infections; driving diverse host outcomes that vary by genotype and environment. *ISME J.* 13, 2068–2081. doi: 10.1038/s41396-019-0419-7

King, J. D., Berry, S., Clarke, B. R., Morris, R. J., and Whitfield, C. (2014). Lipopolysaccharide O antigen size distribution is determined by a chain extension complex of variable stoichiometry in *Escherichia coli* O9a. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6407–6412. doi: 10.1073/pnas.1400814111

Knirel, Y. A., Paramonov, N. A., Shashkov, A. S., Kochetkov, N. K., Yarullin, R. G., Farber, S. M., et al. (1992). Structure of the polysaccharide chains of *Pseudomonas pseudomallei* lipopolysaccharides. *Carbohydr. Res.* 233, 185–193. doi: 10.1016/s0008-6215(00)90930-3

Koonpaew, S., Ubol, M. N., Sirisinha, S., White, N. J., and Chaiyaroj, S. C. (2000). Genome fingerprinting by pulsed-field gel electrophoresis of isolates of *Burkholderia pseudomallei* from patients with melioidosis in Thailand. *Acta Trop.* 74, 187–191. doi: 10.1016/s0001-706x(99)00069-8

Koosakulnirand, S., Phokrai, P., Jenjaroen, K., Roberts, R. A., Utaisincharoen, P., Dunachie, S. J., et al. (2018). Immune response to recombinant *Burkholderia pseudomallei* FliC. *PLoS One* 13:e0198906. doi: 10.1371/journal.pone.0198906

Kronsteiner, B., Chaichana, P., Sumonwiriya, M., Jenjaroen, K., Chowdhury, F. R., Chumseng, S., et al. (2019). Diabetes alters immune response patterns to acute melioidosis in humans. *Eur. J. Immunol.* 49, 1092–1106. doi: 10.1002/eji.201848037

Lazar Adler, N., Stevens, J., Stevens, M., and Galyov, E. (2011). Autotransporters and their role in the virulence of *Burkholderia pseudomallei* and *Burkholderia mallei*. *Front. Microbiol.* 2:151. doi: 10.3389/fmicb.2011.00151

Lazar Adler, N. R., Stevens, M. P., Dean, R. E., Saint, R. J., Pankhania, D., Prior, J. L., et al. (2015). Systematic mutagenesis of genes encoding predicted autotransported proteins of *Burkholderia pseudomallei* identifies factors mediating virulence in mice, net intracellular replication and a novel protein conferring serum resistance. *PLoS One* 10:e0121271. doi: 10.1371/journal.pone.0121271

Lees, J. A., Croucher, N. J., Goldblatt, D., Nosten, F., Parkhill, J., Turner, C., et al. (2017). Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife* 6:e26255.

Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., et al. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* 29, 304–316. doi: 10.1101/gr.241455.118

Lennings, J., West, T. E., and Schwarz, S. (2019). The *Burkholderia* type VI secretion system 5: composition, regulation and role in virulence. *Front. Microbiol.* 9:3339. doi: 10.3389/fmicb.2018.03339

Lieberman, T. D., Flett, K. B., Yelin, I., Martin, T. R., McAdam, A. J., Priebe, G. P., et al. (2014). Genetic variation of a bacterial pathogen within individuals with

cystic fibrosis provides a record of selective pressures. *Nat. Genet.* 46, 82–87. doi: 10.1038/ng.2848

Limmathurotsakul, D., Holden, M. T. G., Coupland, P., Price, E. P., Chantratita, N., Wuthiekanun, V., et al. (2014a). Microevolution of *Burkholderia pseudomallei* during an acute infection. *J. Clin. Microbiol.* 52, 3418–3421. doi: 10.1128/jcm.01219-14

Limmathurotsakul, D., Paeyao, A., Wongratanacheewin, S., Saiprom, N., Takpho, N., Thaipadungpanit, J., et al. (2014b). Role of *Burkholderia pseudomallei* biofilm formation and lipopolysaccharide in relapse of melioidosis. *Clin. Microbiol. Infect.* 20, O854–O856.

Limmathurotsakul, D., Wongratanacheewin, S., Teerawattanasook, N., Wongsuvan, G., Chaisuksant, S., Chetchotisakd, P., et al. (2010a). Increasing incidence of human melioidosis in Northeast Thailand. *Am. J. Trop. Med. Hyg.* 82, 1113–1117. doi: 10.4269/ajtmh.2010.10-0038

Limmathurotsakul, D., Wuthiekanun, V., Chantratita, N., Wongsuvan, G., Amornchai, P., Day, N. P. J., et al. (2010b). *Burkholderia pseudomallei* is spatially distributed in soil in Northeast Thailand. *PLoS Neg. Trop. Dis.* 4:e694. doi: 10.1371/journal.pntd.0000694

Losada, L., Ronning, C. M., DeShazer, D., Woods, D., Fedorova, N., Kim, H. S., et al. (2010). Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biol. Evol.* 2, 102–116. doi: 10.1093/gbe/evq003

Mahasirimongkol, S., Yanai, H., Mushiroda, T., Promphittayarat, W., Wattanapokayakit, S., Phromjai, J., et al. (2012). Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *J. Hum. Genet.* 57, 363–367. doi: 10.1038/jhg.2012.35

Malaria Genomic Epidemiology Network (2008). A global network for investigating the genomic epidemiology of malaria. *Nature* 456, 732–737. doi: 10.1038/nature07632

Manivanh, L., Pierret, A., Rattanavong, S., Kounnavongsa, O., Buisson, Y., Elliott, I., et al. (2017). *Burkholderia pseudomallei* in a lowland rice paddy: seasonal changes and influence of soil depth and physico-chemical properties. *Sci. Rep.* 7:3031.

McRobb, E., Sarovich, D. S., Price, E. P., Kaestli, M., Mayo, M., Keim, P., et al. (2015). Tracing melioidosis back to the source: using whole-genome sequencing to investigate an outbreak originating from a contaminated domestic water supply. *J. Clin. Microbiol.* 53, 1144–1148. doi: 10.1128/jcm.03453-14

Moore, R. A., Tuanyok, A., and Woods, D. E. (2008). Survival of *Burkholderia pseudomallei* in water. *BMC Res. Notes* 1:11.

Moule, M. G., Spink, N., Willcocks, S., Lim, J., Guerra-Assunção, J. A., Cia, F., et al. (2015). Characterization of new virulence factors involved in the intracellular growth and survival of *Burkholderia pseudomallei*. *Infect. Immun.* 84, 701–710. doi: 10.1128/iai.01102-15

Myers, N. D., Chantratita, N., Berrington, W. R., Chierakul, W., Limmathurotsakul, D., Wuthiekanun, V., et al. (2014). The role of NOD2 in murine and human melioidosis. *J. Immunol.* 192, 300–307. doi: 10.4049/jimmunol.1301436

Nandi, T., Holden, M. T. G., Didelot, X., Mehershahi, K., Boddey, J. A., Beacham, I., et al. (2015). *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res.* 25, 129–141. doi: 10.1101/gr.177543.114

Nandi, T., Ong, C., Singh, A. P., Boddey, J., Atkins, T., Sarkar-Tyson, M., et al. (2010). A genomic survey of positive selection in *Burkholderia pseudomallei* provides insights into the evolution of accidental virulence. *PLoS Pathog.* 6:e1000845. doi: 10.1371/journal.ppat.1000845

Nelson, M., Prior, J. L., Lever, M. S., Jones, H. E., Atkins, T. P., and Titball, R. W. (2004). Evaluation of lipopolysaccharide and capsular polysaccharide as subunit vaccines against experimental melioidosis. *J. Med. Microbiol.* 53(Pt 12), 1177–1182. doi: 10.1099/jmm.0.45766-0

Nikolich-Zugich, J., Slifka, M. K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* 4, 123–132. doi: 10.1038/nri1292

Noinarin, P., Chareonsudjai, P., Wangsomnuk, P., Wongratanacheewin, S., and Chareonsudjai, S. (2016). environmental free-living amoebae isolated from soil in khon kaen, Thailand, Antagonize *Burkholderia pseudomallei*. *PLoS One* 11:e0167355. doi: 10.1371/journal.pone.0167355

Norris, M. H., Schweizer, H. P., and Tuanyok, A. (2017). Structural diversity of *Burkholderia pseudomallei* lipopolysaccharides affects innate immune signaling. *PLoS Neg. Trop. Dis.* 11:e0005571. doi: 10.1371/journal.pntd.0005571

Novem, V., Shui, G., Wang, D., Bendt, A. K., Sim, S. H., Liu, Y., et al. (2009). Structural and biological diversity of lipopolysaccharides from *Burkholderia pseudomallei* and *Burkholderia thailandensis*. *Clin. Vaccine Immunol.* 16, 1420–1428. doi: 10.1128/cvi.00472-08

Ooi, W. F., Ong, C., Nandi, T., Kreisberg, J. F., Chua, H. H., Sun, G., et al. (2013). The condition-dependent transcriptional landscape of *Burkholderia pseudomallei*. *PLoS Genet.* 9:e1003795. doi: 10.1371/journal.pgen.1003795

Ou, K., Ong, C., Koh, S. Y., Rodrigues, F., Sim, S. H., Wong, D., et al. (2005). Integrative genomic, transcriptional, and proteomic diversity in natural isolates of the human pathogen *Burkholderia pseudomallei*. *J. Bacteriol.* 187, 4276–4285. doi: 10.1128/jb.187.12.4276-4285.2005

Paksanont, S., Sintiprungrat, K., Yimthin, T., Pumirat, P., Peacock, S. J., and Chantratita, N. (2018). Effect of temperature on *Burkholderia pseudomallei* growth, proteomic changes, motility and resistance to stress environments. *Sci. Rep.* 8:9167.

Panina, E. M., Mattoo, S., Griffith, N., Kozak, N. A., Yuk, M. H., and Miller, J. F. (2005). A genome-wide screen identifies a Bordetella type III secretion effector and candidate effectors in other species. *Mol. Microbiol.* 58, 267–279. doi: 10.1111/j.1365-2958.2005.04823.x

Park, B. S., and Lee, J. O. (2013). Recognition of lipopolysaccharide pattern by TLR4 complexes. *Exp. Mol. Med.* 45:e66. doi: 10.1038/emm.2013.97

Park, J. M., Ghosh, S., and O'Connor, T. J. (2020). Combinatorial selection in amoebal hosts drives the evolution of the human pathogen *Legionella pneumophila*. *Nat. Microbiol.* 5, 599–609. doi: 10.1038/s41564-019-0663-7

Parthasarathy, N., DeShazer, D., England, M., and Waag, D. M. (2006). Polysaccharide microarray technology for the detection of *Burkholderia pseudomallei* and *Burkholderia mallei* antibodies. *Diagn. Microbiol. Infect. Dis.* 56, 329–332. doi: 10.1016/j.diagmicrobio.2006.04.018

Pearson, T., Giffard, P., Beckstrom-Sternberg, S., Auerbach, R., Hornstra, H., Tuanyok, A., et al. (2009). Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. *BMC Biol.* 7:78. doi: 10.1186/1741-7007-7-78

Pearson, T., Sahl, J. W., Hepp, C. M., Handady, K., Hornstra, H., Vazquez, A. J., et al. (2020). Pathogen to commensal? Longitudinal within-host population dynamics, evolution, and adaptation during a chronic >16-year *Burkholderia pseudomallei* infection. *PLoS Pathog.* 16:e1008298. doi: 10.1371/journal.ppat.1008298

Perry, M. B., MacLean, L. L., Schollaardt, T., Bryan, L. E., and Ho, M. (1995). Structural characterization of the lipopolysaccharide O antigens of *Burkholderia pseudomallei*. *Infect. Immun.* 63, 3348–3352. doi: 10.1128/iai.63.9.3348-3352.1995

Pilatz, S., Breitbach, K., Hein, N., Fehlhaber, B., Schulze, J., Brenneke, B., et al. (2006). Identification of *Burkholderia pseudomallei* genes required for the intracellular life cycle and in vivo virulence. *Infect. Immun.* 74, 3576–3586. doi: 10.1128/iai.01262-05

Podnecky, N. L., Rhodes, K. A., Mima, T., Drew, H. R., Chirakul, S., Wuthiekanun, V., et al. (2017). Mechanisms of resistance to folate pathway inhibitors in *Burkholderia pseudomallei* : deviation from the Norm. *mBio* 8:e01357-17.

Pothlichet, J., and Quintana-Murci, L. (2013). The genetics of innate immunity sensors and human disease. *Int. Rev. Immunol.* 32, 157–208. doi: 10.3109/08830185.2013.777064

Price, E. P., Sarovich, D. S., Mayo, M., Tuanyok, A., Drees, K. P., Kaestli, M., et al. (2013). Within-host evolution of *Burkholderia pseudomallei* over a twelve-year chronic carriage infection. *mBio* 4:e00388-13.

Price, E. P., Sarovich, D. S., Smith, E. J., MacHunter, B., Harrington, G., Theobald, V., et al. (2016). Unprecedented melioidosis cases in Northern Australia caused by an asian *Burkholderia pseudomallei* strain identified by using large-scale comparative genomics. *Appl. Environ. Microbiol.* 82, 954–963. doi: 10.1128/aem.03013-15

Price, E. P., Viberg, L. T., Kidd, T. J., Bell, S. C., Currie, B. J., and Sarovich, D. S. (2018). Transcriptomic analysis of longitudinal *Burkholderia pseudomallei* infecting the cystic fibrosis lung. *Microb. Genom.* 4:e000194.

Pumpuang, A., Chantratita, N., Wikraiphat, C., Saiprom, N., Day, N. P., Peacock, S. J., et al. (2011). Survival of *Burkholderia pseudomallei* in distilled water for 16

years. *Trans. R. Soc. Trop. Med. Hyg.* 105, 598–600. doi: 10.1016/j.trstmh.2011.06.004

Rachlin, A., Mayo, M., Webb, J. R., Kleinecke, M., Rigas, V., Harrington, G., et al. (2020). Whole-genome sequencing of *Burkholderia pseudomallei* from an urban melioidosis hot spot reveals a fine-scale population structure and localised spatial clustering in the environment. *Sci. Rep.* 10:5443.

Ralph, A., McBride, J., and Currie, B. J. (2004). Transmission of *Burkholderia pseudomallei* via breast milk in northern Australia. *Pediatr. Infect. Dis. J.* 23, 1169–1171.

Reckseidler, S. L., DeShazer, D., Sokol, P. A., and Woods, D. E. (2001). Detection of bacterial virulence genes by subtractive hybridization: identification of capsular polysaccharide of *Burkholderia pseudomallei* as a major virulence determinant. *Infect. Immun.* 69, 34–44. doi: 10.1128/iai.69.1.34-44.2001

Reckseidler-Zenteno, S. L., DeVinney, R., and Woods, D. E. (2005). The capsular polysaccharide of *Burkholderia pseudomallei* contributes to survival in serum by reducing complement factor C3b deposition. *Infect. Immun.* 73, 1106–1115. doi: 10.1128/iai.73.2.1106-1115.2005

Reckseidler-Zenteno, S. L., Moore, R., and Woods, D. E. (2009). Genetics and function of the capsules of *Burkholderia pseudomallei* and their potential as therapeutic targets. *Mini. Rev. Med. Chem.* 9, 265–271. doi: 10.2174/138955709787316047

Reckseidler-Zenteno, S. L., Viteri, D.-F., Moore, R., Wong, E., Tuanyok, A., and Woods, D. E. (2010). Characterization of the type III capsular polysaccharide produced by *Burkholderia pseudomallei*. *J. Med. Microbiol.* 59(Pt 12), 1403–1414. doi: 10.1099/jmm.0.022202-0

Robinson, W. H. (2015). Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* 11, 171–182. doi: 10.1038/nrrheum.2014.220

Sahl, J. W., Stone, J. K., Gelhaus, H. C., Warren, R. L., Cruttwell, C. J., Funnell, S. G., et al. (2013). Genome Sequence of Burkholderia pseudomallei NCTC 13392. *Genome Announc.* 1:e183-13.

Sahl, J. W., Vazquez, A. J., Hall, C. M., Busch, J. D., Tuanyok, A., Mayo, M., et al. (2016). The effects of signal erosion and core genome reduction on the identification of diagnostic markers. *mBio* 7:9.

Saiprom, N., Sangsri, T., Tandhavanant, S., Sengyee, S., Phunpang, R., Preechanukul, A., et al. (2020). Genomic loss in environmental and isogenic morphotype isolates of *Burkholderia pseudomallei* is associated with intracellular survival and plaque-forming efficiency. *PLoS Neg. Trop. Dis.* 14:e0008590. doi: 10.1371/journal.pntd.0008590

Samatey, F. A., Imada, K., Nagashima, S., Vonderviszt, F., Kumasaka, T., Yamamoto, M., et al. (2001). Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* 410, 331–337. doi: 10.1038/35066504

Sanchez-Buso, L., Comas, I., Jorques, G., and Gonzalez-Candelas, F. (2014). Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat. Genet.* 46, 1205–1211. doi: 10.1038/ng.3114

Sarkar-Tyson, M., Thwaite, J. E., Harding, S. V., Smither, S. J., Oyston, P. C. F., Atkins, T. P., et al. (2007). Polysaccharides and virulence of *Burkholderia pseudomallei*. *J. Med. Microbiol.* 56(Pt 8), 1005–1010.

Sarovich, D. S., Garin, B., Smet, B. D., Kaestli, M., Mayo, M., Vandamme, P., et al. (2016). Phylogenomic analysis reveals an Asian Origin for African *Burkholderia pseudomallei* and further supports melioidosis endemicity in Africa. *mSphere* 1:e00089-15.

Sarovich, D. S., Price, E. P., Webb, J. R., Ward, L. M., Voutsinos, M. Y., Tuanyok, A., et al. (2014). Variable virulence factors in *Burkholderia pseudomallei* (Melioidosis) associated with human disease. *PLoS One* 9:e91682. doi: 10.1371/journal.pone.0091682

Schuppler, M. (2014). How the interaction of Listeria monocytogenes and Acanthamoeba spp. affects growth and distribution of the food borne pathogen. *Appl. Microbiol. Biotechnol.* 98, 2907–2916. doi: 10.1007/s00253-014-5546-5

Scott, A. E., Burtnick, M. N., Stokes, M. G. M., Whelan, A. O., Williamson, E. D., Atkins, T. P., et al. (2014b). *Burkholderia pseudomallei* capsular polysaccharide conjugates provide protection against acute melioidosis. *Infect. Immun.* 82, 3206–3213. doi: 10.1128/iai.01847-14

Scott, A. E., Ngugi, S. A., Laws, T. R., Corser, D., Lonsdale, C. L., D'Elia, R. V., et al. (2014a). Protection against experimental melioidosis following immunisation with a lipopolysaccharide-protein conjugate. *J. Immunol. Res.* 2014: 392170.

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., et al. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66:2888.

Shaw, T., Tellapragada, C., Kamath, A., Kalwaje Eshwara, V., and Mukhopadhyay, C. (2019). Implications of environmental and pathogen-specific determinants on clinical presentations and disease outcome in melioidosis patients. *PLoS Neg. Trop. Dis.* 13:e0007312. doi: 10.1371/journal.pntd.0007312

Shimizu, Y., Ao, H., Soemantri, A., Tiwawech, D., Settheetham-Ishida, W., Kayame, O. W., et al. (2000). Sero- and molecular typing of Duffy blood group in Southeast Asians and Oceanians. *Hum. Biol.* 72, 511–518.

Sidjabat, H. E., Cottrell, K., and Cervin, A. (2015). Draft genome sequences of *Burkholderia pseudomallei* and *Staphylococcus aureus*, Isolated from a Patient with Chronic Rhinosinusitis. *Genome Announc.* 3:e01075-15.

Silverman, J. M., Brunet, Y. R., Cascales, E., and Mougous, J. D. (2012). Structure and regulation of the type VI secretion system. *Annu. Rev. Microbiol.* 66, 453–472. doi: 10.1146/annurev-micro-121809-151619

Sim, S. H., Yu, Y., Lin, C. H., Karuturi, R. K. M., Wuthiekanun, V., Tuanyok, A., et al. (2008). The core and accessory genomes of *Burkholderia pseudomallei*: implications for human melioidosis. *PLoS Pathog.* 4:e1000178. doi: 10.1371/journal.ppat.1000178

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

Sitthidet, C., Stevens, J. M., Chantratita, N., Currie, B. J., Peacock, S. J., Korbsrisate, S., et al. (2008). Prevalence and sequence diversity of a factor required for actin-based motility in natural populations of *Burkholderia* species. *J. Clin. Microbiol.* 46, 2418–2422. doi: 10.1128/jcm.00368-08

Slodkowicz, G., and Goldman, N. (2020). Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5977–5986. doi: 10.1073/pnas.1916786117

Song, L., Yu, Y., Feng, L., He, J., Wang, T., Zhu, H., et al. (2015). Draft genome sequence of *Burkholderia pseudomallei* strain 350105, isolated in Hainan, China, in 1976. *Genome Announc.* 3:e01162-15.

Southeast Asia Infectious Disease Clinical Research Network (2017). Causes and outcomes of sepsis in southeast Asia: a multinational multicentre cross-sectional study. *Lancet Glob. Health* 5, e157–e167.

Speed, D., Holmes, J., and Balding, D. J. (2020). Evaluating and improving heritability models using summary statistics. *Nat. Genet.* 52, 458–462. doi: 10.1038/s41588-020-0600-y

Spring-Pearson, S. M., Stone, J. K., Doyle, A., Allender, C. J., Okinaka, R. T., Mayo, M., et al. (2015). Pangenome analysis of *Burkholderia pseudomallei*: genome evolution preserves gene order despite high recombination rates. *PLoS One* 10:e0140274. doi: 10.1371/journal.pone.0140274

Stevens, J. M., Ulrich, R. L., Taylor, L. A., Wood, M. W., DeShazer, D., Stevens, M. P., et al. (2005). Actin-Binding proteins from *Burkholderia mallei* and *Burkholderia thailandensis* can functionally compensate for the actin-based motility defect of a *Burkholderia pseudomallei* bimA Mutant. *J. Bacteriol.* 187, 7857–7862. doi: 10.1128/jb.187.22.7857-7862.2005

Stevens, M. P., Stevens, J. M., Jeng, R. L., Taylor, L. A., Wood, M. W., Hawes, P., et al. (2005). Identification of a bacterial factor required for actin-based motility of *Burkholderia pseudomallei*. *Mol. Microbiol.* 56, 40–53. doi: 10.1111/j.1365-2958.2004.04528.x

Stone, J. K., Mayo, M., Grasso, S. A., Ginther, J. L., Warrington, S. D., Allender, C. J., et al. (2012). Detection of *Burkholderia pseudomallei* O-antigen serotypes in near-neighbor species. *BMC Microbiol.* 12:250. doi: 10.1186/1471-2180-12-250

Suwannasaen, D., Mahawantung, J., Chaowagul, W., Limmathurotsakul, D., Felgner, P. L., Davies, H., et al. (2011). Human immune responses to *Burkholderia pseudomallei* characterized by protein microarray analysis. *J. Infect. Dis.* 203, 1002–1011. doi: 10.1093/infdis/jiq142

Sweeney, T. E., Perumal, T. M., Henao, R., Nichols, M., Howrylak, J. A., Choi, A. M., et al. (2018). A community approach to mortality prediction in sepsis via gene expression analysis. *Nat. Commun.* 9:694.

Tay, S. T., Cheah, P. C., and Puthucheary, S. D. (2010). Sequence polymorphism and PCR-restriction fragment length polymorphism analysis of the flagellin

gene of *Burkholderia pseudomallei*. *J. Clin. Microbiol.* 48, 1465–1467. doi: 10.1128/jcm.01131-09

Thomas, A. D., Forbes-Faulkner, J., and Parker, M. (1979). Isolation of *Pseudomonas pseudomallei* from clay layers at defined depths. *Am. J. Epidemiol.* 110, 515–521. doi: 10.1093/oxfordjournals.aje.a112832

Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E., et al. (2012). Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* 44, 257–259. doi: 10.1038/ng.1080

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., et al. (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* 489, 443–446. doi: 10.1038/nature11334

Tuanyok, A., Auerbach, R. K., Brettin, T. S., Bruce, D. C., Munk, A. C., Detter, J. C., et al. (2007). A horizontal gene transfer event defines two distinct groups within *Burkholderia pseudomallei* that have dissimilar geographic distributions. *J. Bacteriol.* 189, 9044–9049. doi: 10.1128/jb.01264-07

Tuanyok, A., Leadem, B. R., Auerbach, R. K., Beckstrom-Sternberg, S. M., Beckstrom-Sternberg, J. S., Mayo, M., et al. (2008). Genomic islands from five strains of *Burkholderia pseudomallei*. *BMC Genom.* 9:566. doi: 10.1186/1471-2164-9-566

Tuanyok, A., Stone, J. K., Mayo, M., Kaestli, M., Gruendike, J., Georgia, S., et al. (2012). The Genetic and molecular basis of o-antigenic diversity in *Burkholderia pseudomallei* Lipopolysaccharide. *PLoS Neg. Trop. Dis.* 6:e1453. doi: 10.1371/journal.pntd.0001453

U'Ren, J. M., Schupp, J. M., Pearson, T., Hornstra, H., Friedman, C. L. C., Smith, K. L., et al. (2007). Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. *BMC Microbiol.* 7:23. doi: 10.1186/1471-2180-7-23

van den Broek, T., Borghans, J. A. M., and van Wijk, F. (2018). The full spectrum of human naive T cells. *Nat. Rev. Immunol.* 18, 363–373. doi: 10.1038/s41577-018-0001-y

Vander Broek, C. W., and Stevens, J. M. (2017). Type III secretion in the melioidosis pathogen *Burkholderia pseudomallei*. *Front. Cell. Infect. Microbiol.* 7:255. doi: 10.3389/fcimb.2017.00255

Viberg, L. T., Price, E. P., Kidd, T. J., Bell, S. C., Currie, B. J., and Sarovich, D. S. (2015). Whole-Genome Sequences of five *Burkholderia pseudomallei* isolates from Australian cystic fibrosis patients. *Genome Announc.* 3:e00254-15.

Viberg, L. T., Sarovich, D. S., Kidd, T. J., Geake, J. B., Bell, S. C., Currie, B. J., et al. (2017). Within-host evolution of *Burkholderia pseudomallei* during Chronic Infection of Seven Australasian cystic fibrosis patients. *mBio* 8:e00356-17.

Webb, J. R., Rachlin, A., Rigas, V., Sarovich, D. S., Price, E. P., Kaestli, M., et al. (2019). Tracing the environmental footprint of the *Burkholderia pseudomallei* lipopolysaccharide genotypes in the tropical "Top End" of the Northern Territory, Australia. *PLoS Neg. Trop. Dis.* 13:e0007369. doi: 10.1371/journal.pntd.0007369

Weehuizen, T. A., Prior, J. L., van der Vaart, T. W., Ngugi, S. A., Nepogodiev, S. A., Field, R. A., et al. (2015). Differential toll-like receptor-signalling of *Burkholderia pseudomallei* lipopolysaccharide in murine and human models. *PLoS One* 10:e0145397. doi: 10.1371/journal.pone.0145397

West, T. E., Chantratita, N., Chierakul, W., Limmathurotsakul, D., Wuthiekanun, V., Myers, N. D., et al. (2013). Impaired TLR5 functionality is associated with survival in melioidosis. *J. Immunol.* 190, 3373–3379. doi: 10.4049/jimmunol.1202974

West, T. E., Chierakul, W., Chantratita, N., Limmathurotsakul, D., Wuthiekanun, V., Emond, M. J., et al. (2012). Toll-like receptor 4 region genetic variants are associated with susceptibility to melioidosis. *Genes Immun.* 13, 38–46. doi: 10.1038/gene.2011.49

West, T. E., Ernst, R. K., Jansson-Hutson, M. J., and Skerrett, S. J. (2008). Activation of Toll-like receptors by *Burkholderia pseudomallei*. *BMC Immunol.* 9:46. doi: 10.1186/1471-2172-9-46

Whitlock, G. C., Deeraksa, A., Qazi, O., Judy, B. M., Taylor, K., Propst, K. L., et al. (2010). Protective response to subunit vaccination against intranasal *Burkholderia mallei* and *B. pseudomallei* challenge. *Proc. Vaccinol.* 2, 73–77. doi: 10.1016/j.provac.2010.03.013

Wiersinga, W. J., Wieland, C. W., Dessing, M. C., Chantratita, N., Cheng, A. C., Limmathurotsakul, D., et al. (2007). Toll-like receptor 2 impairs host defense in gram-negative sepsis caused by *Burkholderia pseudomallei* (Melioidosis). *PLoS Med.* 4:e248. doi: 10.1371/journal.pmed.0040248

Wikraiphat, C., Charoensap, J., Utaisincharoen, P., Wongratanacheewin, S., Taweechaisupapong, S., Woods, D. E., et al. (2009). Comparative in vivo and in vitro analyses of putative virulence factors of *Burkholderia pseudomallei* using lipopolysaccharide, capsule and flagellin mutants. *FEMS Immunol. Med. Microbiol.* 56, 253–259. doi: 10.1111/j.1574-695x.2009.00574.x

Willcocks, S. J., Denman, C. C., Atkins, H. S., and Wren, B. W. (2016). Intracellular replication of the well-armed pathogen *Burkholderia pseudomallei*. *Curr. Opin. Microbiol.* 29, 94–103. doi: 10.1016/j.mib.2015.11.007

Williams, N. L., Morris, J. L., Rush, C. M., and Ketheesan, N. (2015). Plasmacytoid dendritic cell bactericidal activity against *Burkholderia pseudomallei*. *Microb. Infect.* 17, 311–316. doi: 10.1016/j.micinf.2014.12.007

Williams, R. C., Muller, Y. L., Hanson, R. L., Knowler, W. C., Mason, C. C., Bian, L., et al. (2011). HLA-DRB1 reduces the risk of type 2 diabetes mellitus by increased insulin secretion. *Diabetologia* 54, 1684–1692. doi: 10.1007/s00125-011-2122-8

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yu, Y., Kim, H. S., Chua, H. H., Lin, C. H., Sim, S. H., Lin, D., et al. (2006). Genomic patterns of pathogen evolution revealed by comparison of *Burkholderia pseudomallei*, the causative agent of melioidosis, to avirulent *Burkholderia thailandensis*. *BMC Microbiol.* 6:46. doi: 10.1186/1471-2180-6-46

Zhang, X., Zhuchenko, O., Kuspa, A., and Soldati, T. (2016). Social amoebae trap and kill bacteria by casting DNA nets. *Nat. Commun.* 7:10938.

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.-J., Butterworth, A. S., Howson, J. M. M., et al. (2017). Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* 49, 1450–1457.

Zou, N., Tsai, S., Feng, S. H., Newsome, T., Kim, H. Y., Li, B., et al. (2008). Relationship between antigenicity and pathogenicity for *Burkholderia pseudomallei* and *Burkholderia mallei* revealed by a large panel of mouse MAbs. *Hybridoma (Larchmt)* 27, 231–240.

Zou, S., Li, J., Zhou, H., Frech, C., Jiang, X., Chu, J. S., et al. (2014). Mutational landscape of intrahepatic cholangiocarcinoma. *Nat. Commun.* 5:5696.

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.

Copyright © 2021 Chomkatekaew, Boonklang, Sangphukieo and Chewapreecha. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/fmicb.2020.619542

Transcriptomic Basis of Serum Resistance and Virulence Related Traits in XDR *P. aeruginosa* Evolved Under Antibiotic Pressure in a Morbidostat Device

Mumina Javed[1,2], Benedikt Jentzsch[1,2], Maximilian Heinrich[1,2], Viola Ueltzhoeffer[1], Silke Peter[1,2], Ulrich Schoppmeier[1], Angel Angelov[3], Sandra Schwarz[1] and Matthias Willmann[1,2,4]**

[1] Interfaculty Institute of Microbiology and Infection Medicine Tübingen, Institute of Medical Microbiology and Hygiene, Tübingen, Germany, [2] German Center for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany, [3] NGS Competence Center Tübingen (NCCT), Tübingen, Germany, [4] Eurofins MVZ Medizinisches Labor Gelsenkirchen, Gelsenkirchen, Germany

Colistin is a last resort antibiotic against the critical status pathogen *Pseudomonas aeruginosa*. Virulence and related traits such as biofilm formation and serum resistance after exposure to sub-inhibitory levels of colistin have been underexplored. We cultivated *P. aeruginosa* in a semi-automated morbidostat device with colistin, metronidazole and a combination of the two antibiotics for 21 days, and completed RNA-Seq to uncover the transcriptional changes over time. Strains became resistant to colistin within this time period. Colistin-resistant strains show significantly increased biofilm formation: the cell density in biofilm increases under exposure to colistin, while the addition of metronidazole can remove this effect. After 7 days of colistin exposure, strains develop an ability to grow in serum, suggesting that colistin drives bacterial modifications conferring a protective effect from serum complement factors. Of note, strains exposed to colistin showed a decrease in virulence, when measured using the *Galleria mellonella* infection model. These phenotypic changes were characterized by a series of differential gene expression changes, particularly those related to LPS modifications, spermidine synthesis (*via speH* and *speE*) and the major stress response regulator *rpoS*. Our results suggest a clinically important bacterial evolution under sub-lethal antibiotic concentration leading to potential for significant changes in the clinical course of infection.

Keywords: multi-drug resistance, biofilm formation, evolutionary trajectories, serum susceptibility, colistin, combination drug therapy, bacterial fitness, automated devices

OPEN ACCESS

Edited by:
Zhan Zhou,
Zhejiang University, China

Reviewed by:
Kwan Soo Ko,
Sungkyunkwan University, South Korea
Oana Ciofu,
University of Copenhagen, Denmark

****Correspondence:***
Mumina Javed
mjaved1@outlook.com
Matthias Willmann
will80@gmx.de

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

Received: *20 October 2020*
Accepted: *30 December 2020*
Published: *25 January 2021*

Citation:
Javed M, Jentzsch B, Heinrich M, Ueltzhoeffer V, Peter S, Schoppmeier U, Angelov A, Schwarz S and Willmann M (2021) Transcriptomic Basis of Serum Resistance and Virulence Related Traits in XDR P. aeruginosa Evolved Under Antibiotic Pressure in a Morbidostat Device. Front. Microbiol. 11:619542. doi: 10.3389/fmicb.2020.619542

INTRODUCTION

Pseudomonas aeruginosa is a gram-negative, opportunistic bacterium and a frequent cause of healthcare acquired infections (HAIs). It belongs to the group of ESKAPE pathogens, which consist of six microorganisms, namely *Enterococcus faecium, Staphylococcus aureus*, *Klebsiella pneumoniae, Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* spp., with a high tendency for causing challenging, drug-resistant, nosocomial infections (Rice, 2008). *P. aeruginosa* accounts for up to 15.4% of bloodstream infections (BSIs) in intensive care units (ICUs) in Western Europe [Antimicrobial resistance and healthcare-associated infections – Annual epidemiological report 2014 (2012 data)], causing high mortality rates.

One of the drugs of last resort against multi drug resistant (MDR) strains of *P. aeruginosa* is colistin. Colistin has a bactericidal effect: as a cationic cyclic peptide, it is able to bind to anionic lipopolysaccharide (LPS) modules and displace $Ca2^{+}$ and $Mg2^{+}$ from the outer cell membrane of *P. aeruginosa*, leading to disruption in the permeability of the membrane, leaking of cell contents, and cell death (Falagas et al., 2005). Although the spread of colistin resistance is not critical to date, resistance has emerged in some instances worldwide, particularly with the increased reliance on colistin for treating multidrug-resistant gram-negative bacterial infections (Lee et al., 2016). It is crucial that new measures are taken to prevent the development of resistance against this last resort drug, including using microbial evolution experiments to uncover the molecular basis of adaptive evolution. Antibiotics are often given in combination in standard therapeutic regimens, with metronidazole being a common drug partner for the treatment of infections by obligate and facultative anaerobic bacteria. It is effective for the management of intra-abdominal infections, gynecological infections, septicemia, endocarditis, bone, and joint infections, amongst several other types of infections (Löfmark et al., 2010). Metronidazole in treatment inhibits DNA synthesis and DNA damage by oxidation, causing single-strand and double-strand breaks that lead to DNA degradation and cell death. Metronidazole is activated when reduced, with molecules binding non-specifically to bacterial DNA, inactivating the DNA and key enzymes of the pathogen; leading to a high level of DNA breakage (Land and Johnson, 1999). An interesting study found that despite metronidazole having no bactericidal effect on *P. aeruginosa*, *in vitro* exposure to a therapeutic concentration of metronidazole increased the number of mutations through induction of the SOS response, thus leading to emergence of antibiotic resistant bacteria (Hocquet and Bertrand, 2014).

Advances in next-generation sequencing of RNA have enabled the analysis of transcriptional changes that occur in bacteria when continuously grown *in vitro* in presence of sub-lethal doses of antibiotics (McPhee et al., 2003; Hood et al., 2013). These evolutionary studies have been extremely valuable in identifying genes and pathways that confer antibiotic resistance, in cases where this is mediated by chromosomal mutations. However, there is a significant lack of mass or high throughput evolutionary studies. In addition, how the transcriptional changes under antibiotic pressure and combination regimens affect other clinically relevant phenotypes remains poorly explored. Such changes may alter the manifestation and severity of the infection as well as the response to treatment and involve factors such as biofilm formation, immune response evasion through serum resistance and virulence. One study which used single drug treatment and examined populations of *P. aeruginosa* evolved in the presence of sublethal concentrations of ciprofloxacin found significant phenotypic changes such as reduced protease activity and swimming motility, as well as increased levels of quorum-sensing (QS) signal molecules (Wassermann et al., 2016). Their results suggested that evolution in the presence of sublethal concentrations of antibiotics have pleiotropic effects on the phenotypes of pathogens. This may promote persistence of the resistant bacterial populations.

The ability of *P. aeruginosa* to form biofilms is associated with severe infections and significant morbidity and mortality (Mulcahy et al., 2014). Biofilms provide *P. aeruginosa* an enormous advantage by promoting survival on medical devices such as catheters, evasion from the immune system, and tolerance to antimicrobial therapy. Increased biofilm formation has also been found to contribute to increased virulence (Ghadaksaz et al., 2015; Maurice et al., 2018). Other virulence factors of *P. aeruginosa* include, amongst others, toxins, exoproteases, phospholipases, the presence of type IV pili and flagella (Ballok and O'Toole, 2013). Serum contains more than thirty proteins of the complement system, and is a crucial component of the host innate immune response which can also initiate the adaptive response. The ability to inhibit complement activation is also considered a virulence trait. However, besides few examples of pathogenic bacteria that carry "*ad hoc*" molecules to block the activity of the complement (Kunert et al., 2007; Miajlovic and Smith, 2014), very little is known about the molecular basis of increased serum resistance and the clinical ramifications overall in *P. aeruginosa*.

We performed the present work using a morbidostat. A morbidostat is a semi-automated culturing device that continuously monitors bacterial growth and adjusts antibiotic concentration to induce bacterial resistance against the drug. In combination with "omic" approaches, experiments using a morbidostat can shed light on the different evolutionary trajectories of antibiotic resistance development as well as further phenotypic modification of clinical relevance. Our previous work with this device concluded that selection for colistin resistance results in a rise of mutations in *pmrB* and *pmrE* in the clinical strain *P. aeruginosa* PA77 (Dößelmann et al., 2017), which are common in other clinical isolates. Here we used the same clinical strain, derived from bloodstream infections and cultured it in our morbidostat for 21 days, under three experimental conditions: no antibiotic, single drug (colistin or metronidazole), and a combination of the two antibiotics. At four key time-points we measured the impact of antibiotic exposure on clinically relevant phenotypes such as antibiotic resistance,

biofilm formation, serum resistance and virulence. We expect that the addition of metronidazole would lead to an alternate transcriptomic and phenotypic profile when compared to single drug treatment isolates. This is the first study to examine the evolution of an extensively drug resistant (XDR) *P. aeruginosa* strain in a morbidostat device under strong antibiotic pressure, and to demonstrate that the acquisition of colistin tolerance and resistance can affect phenotypic traits generally associated with virulence.

MATERIALS AND METHODS

Strain Selection: Characteristics of PA77

The *P. aeruginosa* clinical strain PA77 was isolated at the University Hospital Tübingen, Germany (Willmann et al., 2014). PA77 is extensively drug resistant (XDR) (Magiorakos et al., 2012), being non-susceptible to all antibiotics except colistin and fosfomycin. Multilocus sequence typing analysis show PA77 belongs to the high-risk sequence type (ST) ST308 (Willmann et al., 2014).

Study Design and Experimental Conditions

The morbidostat system was built following the detailed instructions by Toprak et al. (2012) with the modifications outlined in previous work (Dößelmann et al., 2017). As part of our hygiene protocol, we ran 80% ethanol through the tubing system and 3% sodium hypochlorite for 30 min each, and distilled water was used to rinse the tubing after each solution. The biological waste from the morbidostat was fed into a container with a neutralizing solution.

Figure 1 demonstrates an overview of the study design. Strain PA77 was continuously cultivated for 21 days in the morbidostat in four conditions. The first condition was with colistin as a single drug starting with 2 mg/L and with a final concentration of 500 mg/L. The second single drug condition is with 50 mg/L of metronidazole, and the third condition is with a combination of increasing colistin and 50 mg/L metronidazole. The fourth condition was completed as a control, with plain LB medium and no antibiotics. For every condition we ran three replicates in different vials in the morbidostat to investigate whether evolutionary trajectories were stable in all strains. We took samples of the culture at 7, 14, and 21 days of drug exposure, which means collectively there were three replicates over three time points, in four conditions. This totals 37 strains including the baseline strain of PA77 being the original clinical strain not evolved in the morbidostat. All strains were assessed in phenotypic and transcriptomics assays. These samples were processed with RNA*later* Stabilization Reagent (Qiagen, Hilden, Germany) and ultimately frozen *via* the Microbank system (Pro-Lab Diagnostics Inc., TX, United States) at −80°C. All phenotypic assays were performed with one bead taken from these frozen stocks and grown overnight on blood agar plates containing 2 or 16 mg/L colistin. This was to ensure a truly resistant population, devoid of persister cells or dormant bacteria.

DNA Extraction, Library Preparation and Genome Sequencing for Baseline Strain PA77

The genome sequencing was performed at the NGS Competence Center Tübingen (NCCT) using both short (Illumina) and long Oxford Nanopore Technology (ONT) reads for the baseline strain PA77. For ONT, genomic DNA was isolated using Qiagen Genomic Tip G/20, the integrity and quality of the preparations were evaluated on a Femto Pulse capillary electrophoresis device (Agilent, Santa Clara, CA, United States). Library preparation was performed with Ligation Sequencing Kit (SQK-LSK109) with Native Barcoding Expansion (NBD104). The pooled libraries were sequenced on a MinION flow cell (FLO-MIN106D) for 48 h. For short read sequencing, the DNA was extracted with the DNeasy UltraClean Microbial Kit (Qiagen). Illumina sequencing libraries were prepared using Nextera DNA Flex library prep kit (Illumina), using combinatorial dual indexing. The Illumina libraries were pooled and sequenced on a NextSeq 500 High-Output flow cell with 2 × 150 bp.

De novo Genome Assembly and Annotation for Baseline Strain PA77

Whole genome assembly of the baseline strain was performed using the hybrid mode of Unicycler v.048 (Wick et al., 2017). Standard Unicycler parameters were used.

*RNA*Later Storage

Samples were taken from the morbidostat every 2–3 days and stored in *RNA*Later RNA Stabilization Reagent (Qiagen, Hilden, Germany) to preserve the integrity and genetic profile of the isolates. Up to 10^8 cells were centrifuged at 12,000 × *g* for 2 min. The supernatant was removed and the pellet resuspended in 1 ml of *RNA*Later solution and incubated for 1 h at room temperature. The pellet was centrifuged once more at 12,000 × *g* for 5 min, the supernatant removed and the remaining pellet was frozen at −80°C.

RNA Sequencing

Total RNA was extracted with the Quick RNA Fungal/Bacterial Kit (Zymo Research), using lysozyme for cell lysis (0.4 mg/ml). The integrity and quality of the RNA preparations was evaluated with agarose gel electrophoresis and Agilent Bioanalyzer 2100 assays. Total RNAseq Library Kit (Zymo Research) was used for both ribosomal RNA depletion and RNAseq library preparation, using 500 ng total RNA as input. We constructed three libraries per RNA sample. The libraries were sequenced on a NextSeq 500 High-Output flow cell (75 cycles, single reads).

Colistin Susceptibility of PA77 and Isolates Cultivated in Morbidostat

After comparison against broth microdilution (BMD) and other commercial tests with positive results (Javed et al., 2018), antibiotic susceptibility testing was determined using MICRONAUT MIC-Strip (MERLIN Diagnostika Gmbh, Bornheim, Germany): a commercial broth microdilution system

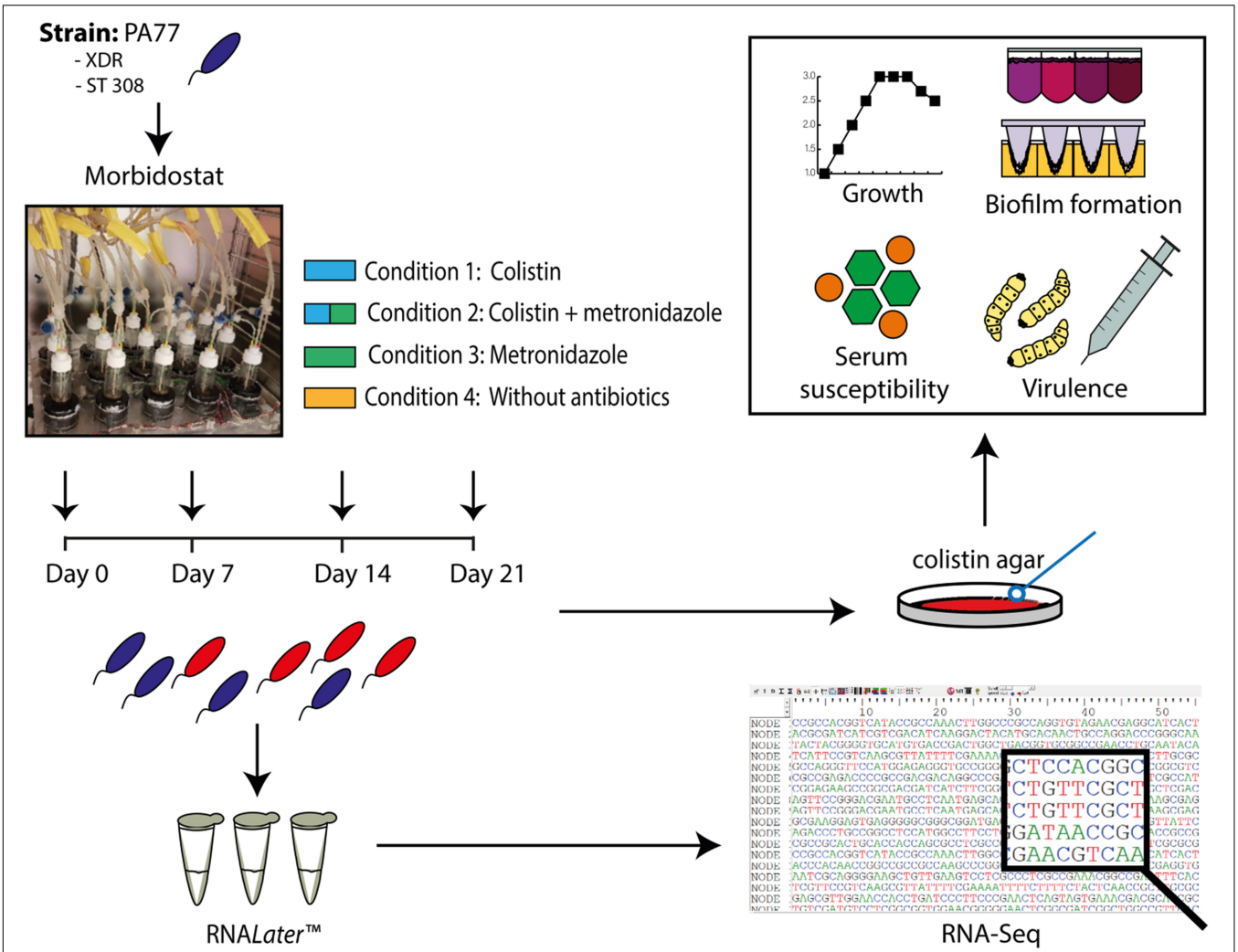


FIGURE 1 | Experimental set up presented as a flow chart. The XDR strain PA77, ST 308, was cultivated in the morbidostat in 20 ml LB medium under four drug conditions: 2–500 mg/L colistin only (Condition 1), a combination of 2–500 mg/L colistin and 50 mg/L metronidazole (Condition 2) and 50 mg/L metronidazole only (Condition 3). A control condition without antibiotics was also completed (Condition 4). Samples of the isolates were taken every 2–3 days, with Day 0, 7, 14, and 21 selected as the key time-points of interest for our study. Isolates generated from the morbidostat were stored in RNA*Later* Stabilization Reagent (Qiagen, Hilden, Germany) to preserve the integrity and genetic profile of the bacteria. Finally, 36 isolates were generated in culture: from four different drug conditions, with three isolates per condition and three experimental time-points. The baseline strain PA77 is considered the Day 0 isolate. XDR, extensively drug resistant; ST, sequencing type. Blue isolates: susceptible to colistin (≤2 mg/L). Red isolates: colistin-resistant (>2 mg/L).

using the international reference methodology (ISO 20776-1) and completed according to the manufacturer's instructions. Briefly, isolates were taken from −80°C stocks and grown on plain agar plates overnight. An inoculum of 0.5 McFarland was prepared in 5 ml NaCl, and then further diluted 1:200 in Mueller Hinton broth. 100 μl of this dilution was pipetted into MIC-Strips which contained freeze dried colistin ranging from 0.0625 to 64 mg/L. The strips were incubated for 18–24 h at 37°C, and results read visually.

Growth Assays

Isolates were grown overnight on selective colistin blood agar plates containing 2 or 16 mg/L colistin. A subculture was made to optical density ($OD_{600\ nm}$) 0.1 in LB medium. 200 μl of this subculture was transferred to sterile, flat-bottomed, polystyrene 96-well microtiter plates. The growth rate of bacteria was determined by incubating the plate at 37°C, measuring the OD of the culture in each well at 600 nm at 30 min intervals for 24 h using a microtiter plate reader (SPECTRAmaxPLUS384 Molecular Devices Inc., United States), and growth curves generated.

Biofilm Formation Assay With Peg Lid Device

Isolates from −80°C storage were grown overnight on sheep blood agar plates containing 2 or 16 mg/L of colistin. After dilution of this culture to 0.5 McFarland in LB medium, 200 μl

was transferred to all but the negative control wells of a flat-bottom 96-well microtiter plate (Nalgene Nunc International, Rochester, NY, United States) Polystyrene microtiter lids were immersed into this subculture (Nalgene Nunc International, Rochester, NY, United States) and incubated aerobically at 37°C for 2 h, followed by incubation at 37°C for 20–22 h anaerobically with Anaerocult tabs (Merck & Co., NJ, United States). The peg lids were dipped in 0.9% NaCl three times for 10 s each to remove planktonic bacteria, and sonicated in an ultrasonic bath (Sonorex^{TM} RK100) for 10 min. The peg lid is then placed in a mixture of 75 μl CHAPS and 75 μl EDTA solution on a rocking table (20 Hz) for 1 h, and the number of viable CFU per ml was determined by serial dilution plating, and counting colonies after 20–24 h incubation at 37°C.

Biofilm Density Quantification With Crystal Violet Staining

Isolates from −80°C storage were grown overnight on sheep blood agar plates containing 2 or 16 mg/L of colistin. After dilution of this culture to 0.5 McFarland in LB medium, 200 μl was transferred to all but the negative control wells of a flat-bottom 96-well microtiter plate (Nalgene Nunc International, Rochester, NY, United States) and incubated aerobically at 37°C for 24 h. Planktonic cells were stained with an aqueous solution of 0.1% crystal violet for 30 min. The excess crystal violet was discarded, and wells were rinsed with distilled water. Stained biofilms were resuspended in 5% acetic acid for 30 min, and absorbance was measured at 590 nm by a microtiter plate reader. Assays were performed in triplicate in three independent experiments, with standard deviations indicated.

Serum Selection

Normal human serum (NHS) from five healthy donors (Department of Transfusion Medicine, University Hospital Tübingen) was stored in aliquots at −80°C. The optimum serum dilution was determined for the baseline strain PA77. Normal human serum (NHS) from healthy donors (Department of Transfusion Medicine, University Hospital Tübingen) was stored in aliquots at −80°C. Strain PA77 was incubated with a percentage of serum, ranging from 10 to 90%. Serum was diluted with PBS (Gibco, Gaithersburg, MD, United States). They were then incubated with a luciferase compound at 37°C for 0, 2, and 4 h and the luminescence of ATP produced in culture measured as an indication of growth. It was determined that 50% was the optimum concentration of serum to use for this study.

Serum Killing Assay

Serum killing assays were completed with BacTiter-Glo^{TM} Microbial Cell Viability Assay (Promega, Madison, WI, United States) as described (Necchi et al., 2017). Normal human serum (NHS) from healthy donors (Department of Transfusion Medicine, University Hospital Tübingen) was stored in aliquots at −80°C. Heat inactivated serum (HIS) was generated by incubating the serum at 56°C for 30 min.

Overnight culture of bacteria was diluted to OD 600 nm 0.1 and subcultured for 1 h. Strains were incubated at 37°C in 100 μl HIS- or NHS-PBS in a 96 well V-bottom microtiter plate (Greiner bio-one, Frickenhausen, Germany) in triplicates for 0, 2, 4, and 6 h. After incubation, plates were centrifuged at 3,500 × *g* for 5 min and the pelleted bacteria were resuspended in 100 μl PBS. To determine the number of viable bacterial cells, 50 μl bacterial suspension and 50 μl BacTiter-Glo^{TM} reagent were transferred to a white LUMITRAC^{TM} 96 well F-bottom microtiter plate (Greiner bio-one, Frickenhausen, Germany) and the adenosine triphosphate (ATP) levels produced by the bacteria were quantified with a Tecan Infinite® 200 PRO. Resulting luminescence values were log transformed (natural logarithm) and the linear regression coefficients of the resulting growth curves in serum (log luminescence over time) were used to calculate a coefficient difference (CD). Here, the regression coefficient of the growth in HIS (αHIS) was subtracted from the regression coefficient of growth in NHS (αNHS). CD = αNHS – αHIS.

Growth in HIS was considered a growth control, hence the coefficient needed to be positive to be considered an adequate experimental setup. A CD value below 0 indicates either effective killing of bacteria by the serum or growth inhibition by the serum. Serum killing happened when the NHS coefficient was negative. Higher negative values would generally indicate a stronger impact of the serum on the bacteria, lower negative values indicate the development of resistance to serum. CD values $\geq$ 0 indicates no effect of the serum on growth, meaning a resistance to serum.

In vivo Virulence Assays

Galleria mellonella larvae of the same size and weight range were purchased from Biosystems Technology (TruLarv^{TM}). Subcultured bacteria were serially diluted in PBS to 8–10 CFU. Each *G. mellonella* larvae was injected with 10 μl of 8–10 CFU bacterial dilution using a 30 gage syringe (BD Biosciences, Allschwil, Switzerland). The larvae were then incubated at 37°C and monitored for 36 h after infection, and the death events were recorded every 2 h. Death was defined as when the larvae stopped responding to touch. Ten microliter aliquots of the bacterial dilutions injected into the larvae were plated in triplicates on LB agar plate and the CFU was determined after overnight incubation to ensure that the injected inoculum was in the 8–10 CFU/10 μl range. A hazard ratio (HR) was calculated with a Cox proportional hazard model (using Stata version 12.1), factoring in the CFU injected into each larvae.

Identification of Differentially Expressed Genes With Geneious Prime

We used Geneious Prime (version 2020.0.5) to analyze and visualize the transcriptomic sequencing data. Reads from morbidostat-generated isolates were pre-processed with Trimmomatics (Bolger et al., 2014) aligned to the reference sequence PA77 using Bowtie2 (Langmead and Salzberg, 2012). We used six replicates per isolate, with at least 10 mio high quality reads. Expression levels of each isolate were compared, and volcano plots generated with the plugin DESeq2

(Love et al., 2014). The genes of interest were filtered by selecting the ratio of significance to 25 and a log2 fold expression of ±4.

Statistical Analysis

GraphPad PRISM 5 (GraphPad Software, San Diego, CA, United States) and Stata version 12.1 (Stat Corp., College Station, TX, United States) were used to perform null hypothesis testing. All error bars represent standard deviation. For each figure, the number of replicates and other information relevant for assessing the accuracy and precision of the measurements are included in the corresponding legend.

RESULTS

Characteristics of PA77 and Morbidostat Generated Isolates From NGS

Hybrid assembly of the baseline strain PA77 resulted in two contigs with total length of 6.9 Mbp. The baseline strain was used as the template for differential gene expression analysis.

No Changes in Growth Rates Between Baseline, Control and Experimental Isolates

Isolates from each condition and key time points were cultured in plain LB medium and growth measured every 30 min *via* OD reader. The results are displayed as Log-OD increase per hour (**Figure 2**) and as growth curves in **Supplementary Figure 1**. There was no difference in the log increase in OD between isolates, indicating that the growth kinetics of the evolved isolates were similar to those of the baseline strain.

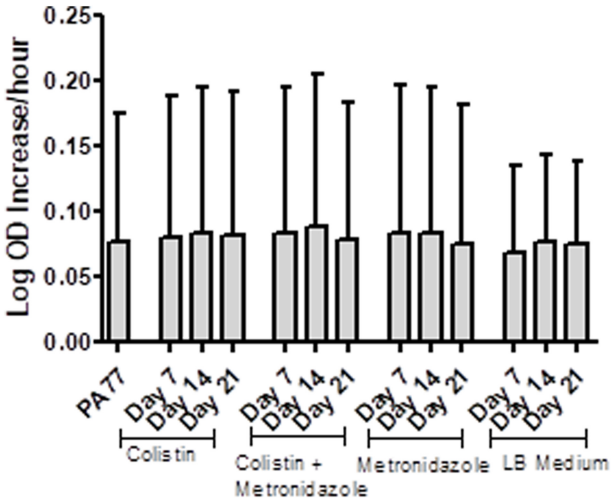


FIGURE 2 | No changes in the growth rates of baseline strain and evolved isolates. Isolates were inoculated in plain LB medium and the optical density measured every 30 min at OD 600 nm, during incubation at 37°C for 24 h, with three replicates per isolate. The results are presented at Log OD increase per hour, with error bars representing the 95% confidence intervals.

Resistance to Colistin Increases as Exposure in Morbidostat Continues for Up to 21 days

PA77 strains continuously exposed to colistin reached a colistin MIC > 64 mg/L at Day 21. This is at least 16-fold higher than the clinical EUCAST breakpoint of colistin for *P. aeruginosa* at ≥2 mg/L (EUCAST, 2020; **Figure 3**). The metronidazole-only and LB medium control strains did not develop colistin resistance. The full MIC values are presented in **Supplementary Table 1**.

Increase in Number of Viable Cells, Density and Biomass in Biofilm After Colistin Exposure

Next, we assessed whether the ability to form biofilms was altered under antibiotic exposure.

Looking at the number of viable cells in biofilm, by far the biggest increase in development of biofilm occurred in the colistin only condition at Day 21, with strains producing 30-fold more viable cells in biofilm than the baseline strain ($p < 0.001$) (**Figure 4A**). All three time-points in this condition showed significantly increased biofilm production relative to the baseline. Looking deeper between drug conditions, we saw an increase in the number of viable cells in biofilm for isolates exposed to colistin for 7 days, with 14.5×10^7 viable cells in biofilm compared to isolates exposed to metronidazole for the same time period with 6.7×10^7 viable cells in biofilm ($p < 0.001$), and in

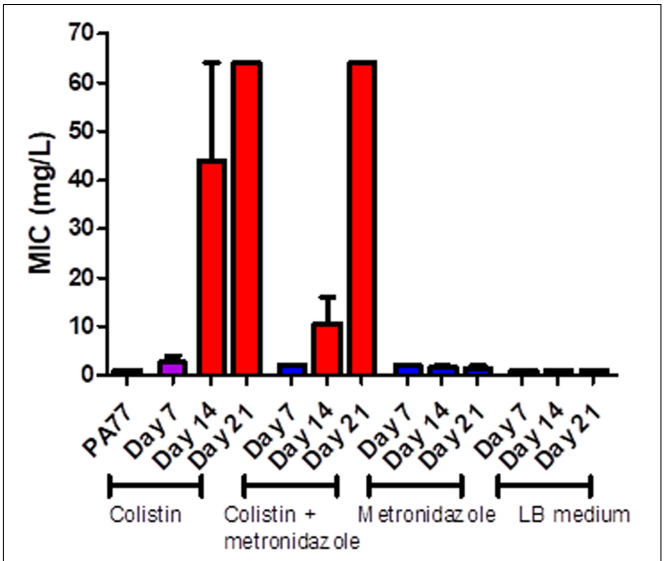


FIGURE 3 | Measure of colistin sensitivity in baseline strain PA77 and evolved isolates. Distribution of the minimum inhibitory concentrations (MICs, mg/L) for colistin ($n = 37$) with isolates cultivated in the morbidostat across four different conditions and samples at three time-points: seven, 14 and 21 days of cultivation. The results are presented as the mean of three biological replicates and range. Blue: indicates susceptibility to colistin (MIC ≤ 2 mg/L). Red: indicates resistance to colistin (MIC > 2 mg/L). Purple: indicates a mix of replicates that are both susceptible and resistance to colistin.

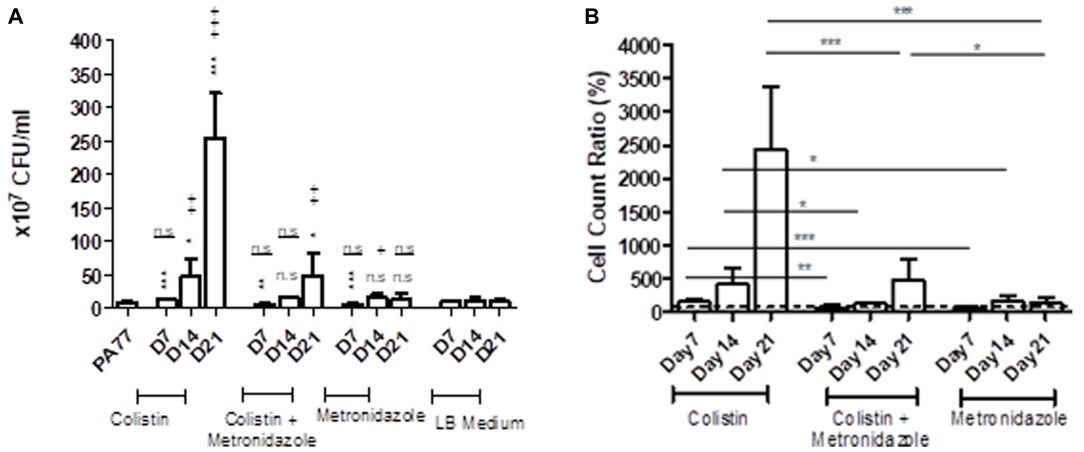


FIGURE 4 | Altered number of viable cells in biofilms of the evolved isolates. **(A)** Measure of viable cells in biofilm produced by isolates cultivated in the morbidostat for 21 days in four different conditions: colistin only, a combination of colistin and metronidazole, metronidazole only and no antibiotic as a control condition. Six replicates per isolate were taken from the morbidostat and grown in LB medium for 24 h and the resulting biofilm detached chemically and mechanically, serially diluted and plated on agar. The CFU was counted and recorded as values from 10^7. The statistical analysis shown are averages of at least three independent experiments ± sd. Comparison to the baseline strain PA77 is indicated by *n.s, not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Comparison to LB medium isolates is denoted using +. n.s, not significant; +$p < 0.05$; ++$p < 0.01$; +++$p < 0.001$. **(B)** The values in panel A were adjusted to a ratio relative to LB medium isolates and displayed here as a bar chart, with a *t*-test used to calculate the difference between the three experimental conditions at each time point. *$p < 0.05$, ***$p < 0.001$. The dashed line represents the values for the control isolates, which is set to 100%.

the combination drug condition (6.1×10^7 viable cells in biofilm) ($p < 0.05$) (**Figure 4B**).

Within 14 days of exposure to colistin as a single drug, there was a faster trajectory of increased biofilm formation than for isolates exposed to metronidazole only ($p < 0.05$) or in the combination of the two drugs ($p < 0.05$) (**Figure 4B**), which may indicate that metronidazole begins to affect the biofilm forming capabilities of the isolates. When compared to the baseline, there were a higher number of viable cells in biofilm in metronidazole-only isolates (17.5×10^7) ($p < 0.001$) at Day 14 than the other conditions at this time-point (**Figure 4A**). After 21 days of exposure, isolates from the combination drug condition had fivefold more viable cells in biofilm than the baseline ($p < 0.01$), while the biofilm formed by metronidazole only isolates did not differ significantly. Between drug conditions, the metronidazole-only condition isolates showed 17-fold decrease in viable cells in biofilm after 21 days of exposure (14.7×10^7 viable cells in biofilm) compared to colistin-only isolates at Day 21 (254.1×10^7) ($p < 0.001$) (**Figure 4B**). The combination drug condition isolates also showed a slower trajectory of increased biofilm formation, with 80.5% less biofilm produced at Day 21 compared to the colistin-only condition ($p < 0.001$) (**Figure 4B**).

The metronidazole-only isolates showed a general increase in biomass in biofilm when compared to both the baseline strain and the control LB medium isolates (**Figure 5A**). The adjusted values showed isolates sampled at Day 7 in all three drug conditions show a decrease in biomass produced in biofilm compared to control isolates, with colistin-only isolates producing 76.2% of biomass in biofilm, and combination isolates producing 66.9% of biomass in biofilm. However, isolates cultivated in metronidazole only show an increase in biomass in biofilm compared to the two other drug conditions (**Figure 5B**). Isolates exposed to metronidazole for 14 days produced 26% more biomass in biofilm than isolates exposed to colistin and metronidazole at the same time point ($p < 0.05$) (**Figure 5B**). Isolates exposed to colistin for 14 days produced 36% less biomass in biofilm than isolates cultivated in metronidazole only ($p < 0.01$), and 10% less biomass in biofilm compared to combination drug isolates ($p < 0.05$). The isolates exposed to colistin and metronidazole for 21 days showed the highest increase in biomass in biofilm compared to the isolates cultured in the two other drug conditions, with 159% biomass measured, compared to 128.3% for colistin isolates ($p < 0.01$), and 124.6% biomass measured for metronidazole only isolates ($p < 0.01$).

Isolates Become Resistant to Complement Factors in Serum Within 7 days of Colistin Exposure

Next, we tested resistance of the evolved isolates to complement factors (**Figure 6** and **Supplementary Figures 2A–M**), with resistance being defined from the ability of these isolates to grow in human serum. The coefficient difference (CD) for the baseline strain is −1.01, which indicates a killing effect: the bacteria are not able to grow in 50% serum. However, after 7 days of exposure to colistin in the morbidostat, the isolates have a CD of −0.07, which indicates development of resistance to complement factors in the same serum. At Day 14 of colistin exposure, isolates were still resistant to serum complement, but less than at Day 7 (CD = −0.129). At Day 21, we see the biggest increase in serum resistance, with a CD of 0.108.

In the combination drug condition, there was also a notable development to serum resistance by Day 7 with a CD of −0.01. At Day 14 and Day 21, strains became again slightly more

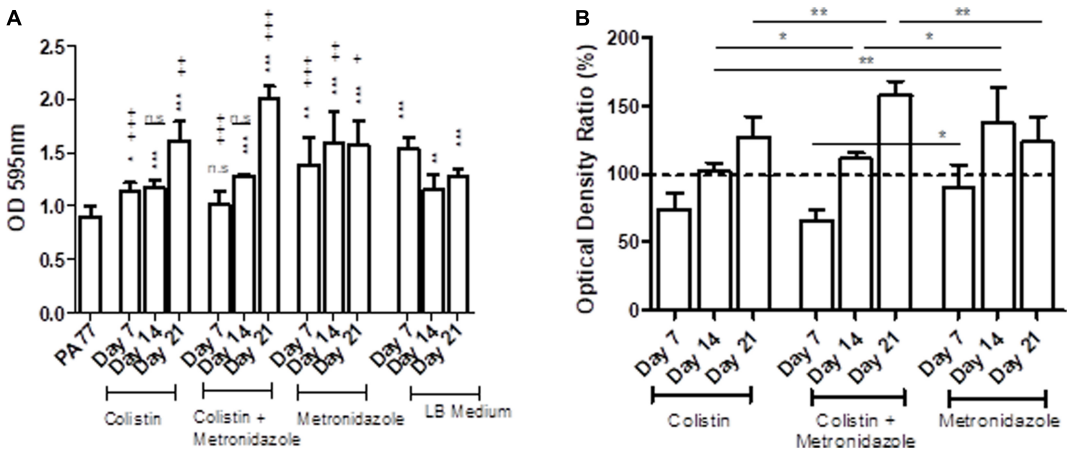


FIGURE 5 | Changes in biomass in biofilms of the evolved isolates. **(A)** The OD values of biofilm staining assays using six replicates per isolate and the standard deviation are presented for strains cultivated in the morbidostat for 21 days in four different conditions: colistin only, a combination of colistin and metronidazole, metronidazole only and in LB medium as a control. The baseline strain PA77 was used as a Day 0 comparison. A *t*-test was used to compare the time-points to each other. Comparison to the baseline strain PA77 is indicated by *n.s, not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Comparison to LB medium isolates is denoted using +. n.s, not significant; $+p < 0.05$; $++p < 0.01$; $+++p < 0.001$. **(B)** The values of the experiments in panel A have been adjusted as a ratio, relative to the values of the LB medium isolates and displayed in a bar chart as percentages. The three experimental conditions have been compared according to the time point, and a *t*-test was used to calculate the significance. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The dashed line represents the values for the control isolates, and is set to 100%.

susceptible to serum (CD −0.148 and −0.139, respectively). The metronidazole-only and control condition do not show a strong deviation from the baseline susceptibility to serum.

Decreased Virulence in *G. mellonella* After Infection With Isolates Exposed to Colistin for 21 days

We examined the virulence of the evolved isolates in a *G. mellonella* model of infection (Tsai et al., 2016; **Figures 7A–C** and **Supplementary Figures 3A–C**). Seven days of single exposure to colistin did not lead to a significant change in virulence (HR = 0.75, lower 95% CI: 0.40, upper 95% CI: 1.13, $p = 0.13$). At Day 14 we still did not observe a significant deviation from baseline. However, at Day 21 we noted a significant attenuation of virulence potential compared to the baseline (HR = 0.36, lower 95% CI: 0.20, upper 95% CI: 0.64, $p = 0.01$). This effect was not seen in the combination drug condition, with metronidazole seemingly having a modulatory effect on the impact of colistin exposure on virulence. Isolates cultured with metronidazole as a single-drug condition also demonstrated no statistically significant changes in virulence. This was the same for isolates in the control condition (LB medium only).

Significant Variation in the Transcriptomic Profile of the Evolved Isolates Compared to Control Isolates

To examine the transcriptional profile of the isolates that had been evolved under different drug pressures, we performed RNA-Seq analysis on *P. aeruginosa* PA77 at the baseline and on strains grown for 7, 14, and 21 days in presence of single or a combination of antibiotics. Due to the drastic changes in the phenotype of the morbidostat-generated strains, namely increased biofilm formation, decrease in virulence and loss of susceptibility to serum, we wanted to investigate general changes in the differentially expressed genes between the three time-points and four conditions (**Figures 8A,B** and **Supplementary Figures 2A–J**). For differential gene expression, a threshold of log2 fold change (FC) of ± 4 was applied.

Among the genes that were upregulated include the *arnBCADTEF* operon (also known as *pmrHFIJKLME*), which has a crucial role in colistin resistance in *P. aeruginosa* through the addition of 4-amino-4-deoxy-L-arabinose (L-Ara4N) to lipid A (Lo Sciuto et al., 2020). The *arnBCADTEF* operon is upregulated in colistin within 21 days (**Figure 8A**), but not differentially expressed at a significant level in isolates exposed to LB medium (**Figure 8B**) and metronidazole only (**Supplementary Figure 2J**).

Two other gene transcripts, *speH* and *speE*, were significantly upregulated in the colistin only and combination drug conditions, as part of the operon *speEH-pmrAB*. The highest positive log fold change for *speE* occurred in colistin Day 21 isolates with 7.25 (differential absolute confidence 128.25, $p < 0.001$) (**Figure 8A**) followed by colistin and combination isolates after 7 days of drug exposure: 6.83 (differential absolute confidence: 59.5, $p < 0.001$) and 5.95 (differential absolute confidence: 56.43, $p < 0.001$), respectively (**Supplementary Table 2**). In contrast, *speE* was downregulated in the medium-control isolates, with a log FC of −0.29 at Day 7 (differential absolute confidence: −1.96, *p* value: 0.006), −0.08 at Day 14 (differential absolute confidence: −0.06, $p < 0.001$) and a log fold change of 1.18 at Day 21 (differential absolute confidence: 5.9, $p < 0.001$) (**Supplementary Table 2**).

One gene of interest that was downregulated is *rpoS*. The highest fold change difference for this gene transcript in particular occurs after 7 days of exposure to metronidazole, with a downregulation in expression at −5.74 (differential absolute

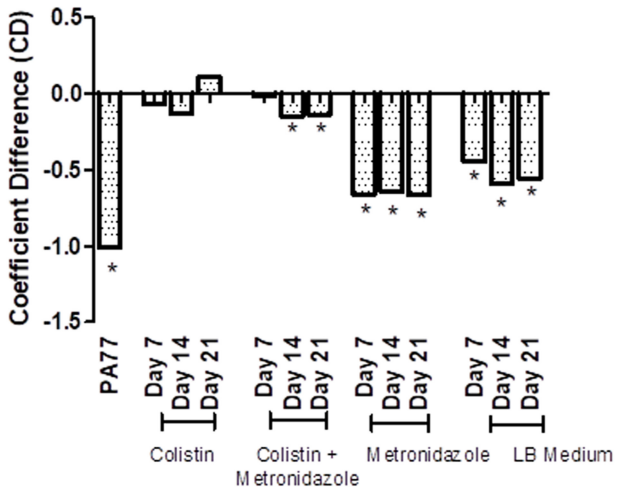


FIGURE 6 | Transient decrease in susceptibility to serum in strains evolved in presence of colistin. Susceptibility to serum was measured by incubating *P. aeruginosa* isolates with 50% human serum and a luciferase compound, and measuring luminescence of ATP produced over 6 h. The results are presented as a bar graph with coefficient difference values calculated using log regression. Isolates with a negative αNHS value, indicating that they are killed in serum, are marked with an asterisk (*). PA77: baseline strain. Colistin: colistin only. Colistin + metronidazole: a combination of colistin and 50 mg/L metronidazole. Metronidazole: 50 mg/L metronidazole. LB medium: control condition isolates where isolates were not exposed to antibiotics and only cultured in LB medium.

confidence −95.99, $p < 0.001$) (**Supplementary Table 2**), closely followed by isolates cultured in colistin for 14 days (log2 FC: −5.25, differential absolute confidence: −129.11, $p < 0.001$) and isolates cultivated in a mix of colistin and metronidazole for 7 days (log2 FC: −5.22, differential absolute confidence: −156.4, $p < 0.001$). Isolates in the medium-control condition fall below the threshold of ≤−4 log2 ratio FC, but are clearly downregulated none the less, ranging from log2 FC of −0.46 (LB medium Day 14) to −2.42 (LB medium Day 21) (**Supplementary Table 2**).

DISCUSSION

In this work, we used one XDR clinical strain of *P. aeruginosa* and cultured them in a morbidostat device with two antibiotics for 21 days, in order to discover the many ways the strain can evolve resistance in a simulated clinical setting. The aim of this work was to investigate the specific contribution of metronidazole alongside the development of colistin resistance in *P. aeruginosa* and evaluate the effect of antibiotic stress and on bacterial fitness, geno- and phenotype.

A number of assays were performed to elucidate the changes in the bacterial phenotype under distinct antibiotic pressure. Subinhibitory exposure of colistin has a clear effect in many of the factors that we have looked at: by the end of 21 days there was increased biofilm formation, a loss in susceptibility to serum and a decrease in virulence. Metronidazole seemingly contributes a modulatory effect on two of these factors: with less viable cells in biofilm and less susceptibility to serum complement factors in combination and metronidazole only condition isolates, as measured by quantifying the concentration of ATP in serum over time. Interestingly, isolates cultured in colistin, both as a single drug and in combination are unable to grow in either heat-inactivated serum or active serum at 6 h, unlike isolates cultivated in metronidazole, which continue to grow in heat-inactivated serum. This indicates a developed sensitivity to heat-stable factors in serum as a result of exposure to sub-inhibitory concentrations of colistin.

Analysis of the transcriptomic profile uncovered a series of differential gene expression patterns. One gene of interest is *speE*, encoding a spermidine synthase, an enzyme that catalyzes the irreversible transfer of a propylamine group from the amino donor *S*-adenosylmethioninamine (decarboxy-AdoMet) to putrescine (1,4-diaminobutane) to yield spermidine. Polyamines such as spermidine are involved in several biological processes, including binding to nucleic acids, stabilizing membranes, and protecting against host immune responses (Chou et al., 2008). Spermidine may be of clinical importance as a biofilm inhibitor: catheters immersed with norspermidine were effective in disrupting mature biofilm (Qu et al., 2016). Exogenous spermidine also protected *P. aeruginosa* against polymyxin B through stabilization of lipopolysaccharides (LPS), while endogenously synthesized spermidine conferred a protective effect against the host immune response to clinical strains of *P. aeruginosa* that produce high amounts of it (Johnson et al., 2012). Its upregulation as part of the operon *speEH-pmrAB* is interesting, as *pmrAB* genes are well established as mediating resistance to colistin and other antibiotics (Lee et al., 2005; Bricio-Moreno et al., 2018) with one study finding that the PmrAB regulon activates *speE* in the presence of antimicrobial peptides such as colistin (Bolard et al., 2019).

Our transcriptomics results indicated that exposure to increasing concentrations of colistin led to significant increased expression of *speH* and *speE* (*speE/H*). The upregulation of these genes may be a mechanism to confer tolerance to colistin under antibiotic pressure in the morbidostat, particularly as it is not highly differentially expressed in isolates exposed to metronidazole only and is found upregulated in isolates that are colistin-resistant. The efficacy of colistin is directly related to the LPS structure, specifically through binding to anionic LPS components of the outer membrane, causing increase of cell permeability and cell death by cell lysis. *P. aeruginosa* defends against colistin by the addition of l-4-aminoarabinose (l-Ara4N) to lipid A phosphates (Nowicki et al., 2014). The proteins for the synthesis and transfer of l-Ara4N are encoded by the *arnBCADTEF* operon and are regulated by the PmrAB and PhoPQ two-component regulatory systems. *In vitro* evolution studies confirm this effect: high-level colistin resistance does not evolve in the absence of a functional *arnBCADTEF* operon (Lo Sciuto et al., 2020).

Upregulation of the *arnBCADTEF* operon–as we have observed it under colistin exposure–results in decreased polymyxins binding to the cell surface and the development of cross-resistance to colistin and other antibiotics (Romano et al., 2019). The LPS is also an inducer of the complement system,

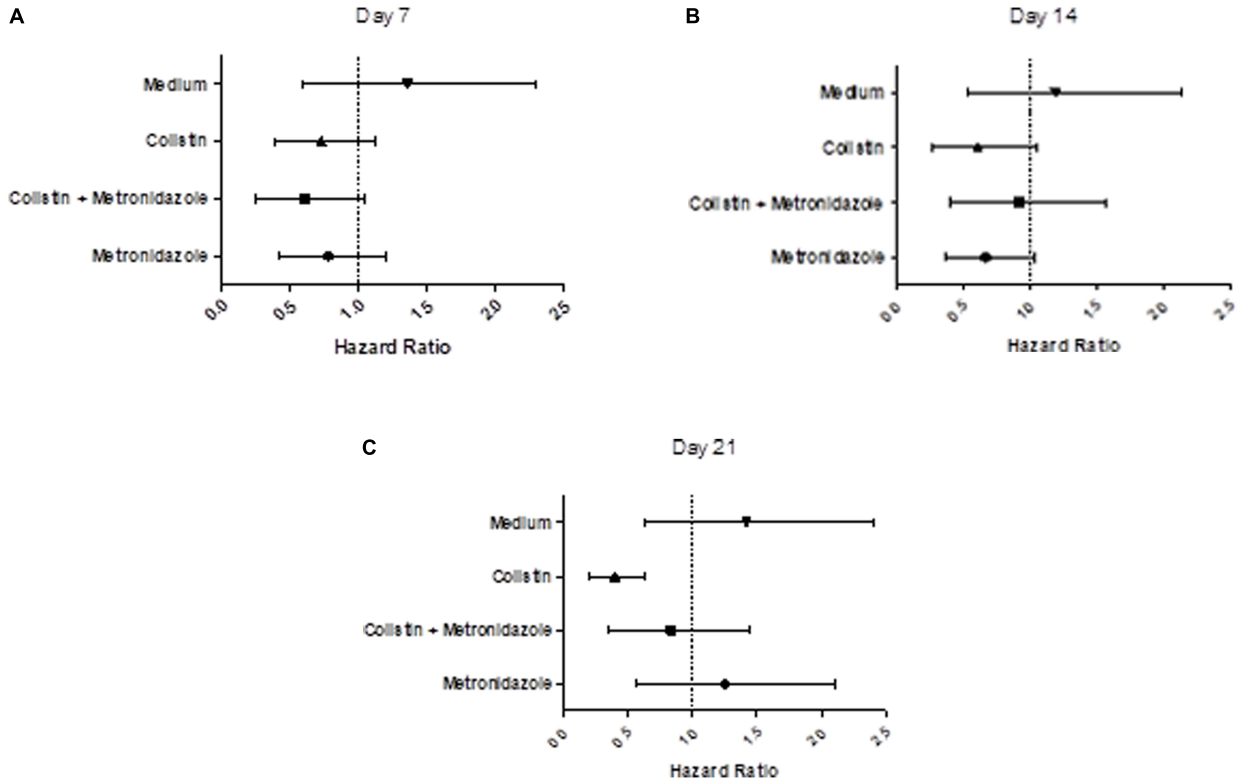


FIGURE 7 | Prolonged exposure to colistin decreases virulence of *P. aeruginosa* PA77 in a *G. mellonella* model of infection. **(A–C)** The virulence potential of isolates was measured with *G. mellonella* larvae. The results are displayed as forest plots per time-point. The hazard ratio is calculated relative to the baseline condition strain PA77. Colistin (▲): colistin only. Colistin + metronidazole (■): a combination of colistin and 50 mg/L metronidazole. Metronidazole (●): 50 mg/L metronidazole. LB medium (▼): control condition isolates not exposed to antibiotics and only cultured in LB medium. **(A)** Day 7: Compared to the baseline strain, there is no significant change in virulence potential of these strains. **(B)** Day 14: There is no significant effect on virulence potential in any of these conditions overall, relative to the baseline strain. **(C)** Day 21: The isolates exposed to colistin in the single-drug condition show a decrease in virulence relative to the baseline strain.

activated *via* its components lipid A, core, and O-antigen. Bacteria expressing long O-antigen chains are usually more resistant to serum than their O antigen-deficient isogenic mutants (Hong and Payne, 1997; Burns and Hull, 1998; Bravo et al., 2008) and are particularly necessary for serum resistance in *P. aeruginosa* (Ohno et al., 1995). Naturally, there was no resistance to colistin for isolates exposed to only metronidazole; these isolates also did not become resistant to serum and the gene transcripts *speE/H* and the *arnBCADTEF* operon were not upregulated significantly (**Supplementary Table 2**). Resistance to complement is strongly associated with the capability of systemic survival, multiplication, and spread of a wide range of Gram-negative pathogens (Short et al., 2020). Therefore we speculate that modifications in the LPS *via* upregulation of the *speE/H* genes and *arnBCADTEF* operon, regulated by pmrAB, may contribute to colistin resistance and loss of susceptibility to serum in our colistin-resistant isolates generated in the morbidostat.

The results of the virulence assays completed using *G. mellonella* show a significant decrease in virulence for one type of isolate: those exposed to colistin for 21 days (**Figure 7C** and **Supplementary Figure 3C**). The remaining isolates over the four conditions and three time-points show no significant difference in virulence potential when compared to the baseline strain PA77. This effect has been found in another recent study using Gram negative bacteria (Esposito et al., 2018). This study that assessed the virulence profile of 16 colistin resistant *K. pneumoniae* isolates with different levels of colistin resistance, found that the colistin MIC of *K. pneumoniae* isolates is predictive of their lethality (LD_{50} and LD_{90} values) in *G. mellonella*. High colistin MIC values were predictive of lower virulence of the isolates, indicating that genetic adaption to high levels of colistin resistance could somehow impair *K. pneumoniae* infectivity, although they did not observe any significant correlation between colistin-resistance mechanisms and virulence. In other research, *P. aeruginosa* with LPS deficiencies display attenuation in virulence in *G. mellonella* (Jarrell' and Kropinski, 1982; Tsai et al., 2016) and so it is likely that isolates exposed to colistin for 21 days may have undergone transcriptional changes related to LPS modifications that make them lose their virulence potential. Metronidazole alone as a single drug treatment did not have any effect on virulence. Isolates exposed to metronidazole in combination with colistin, tend to be relatively virulent when compared to the single drug isolates. This indicates that metronidazole slows down the trajectory of

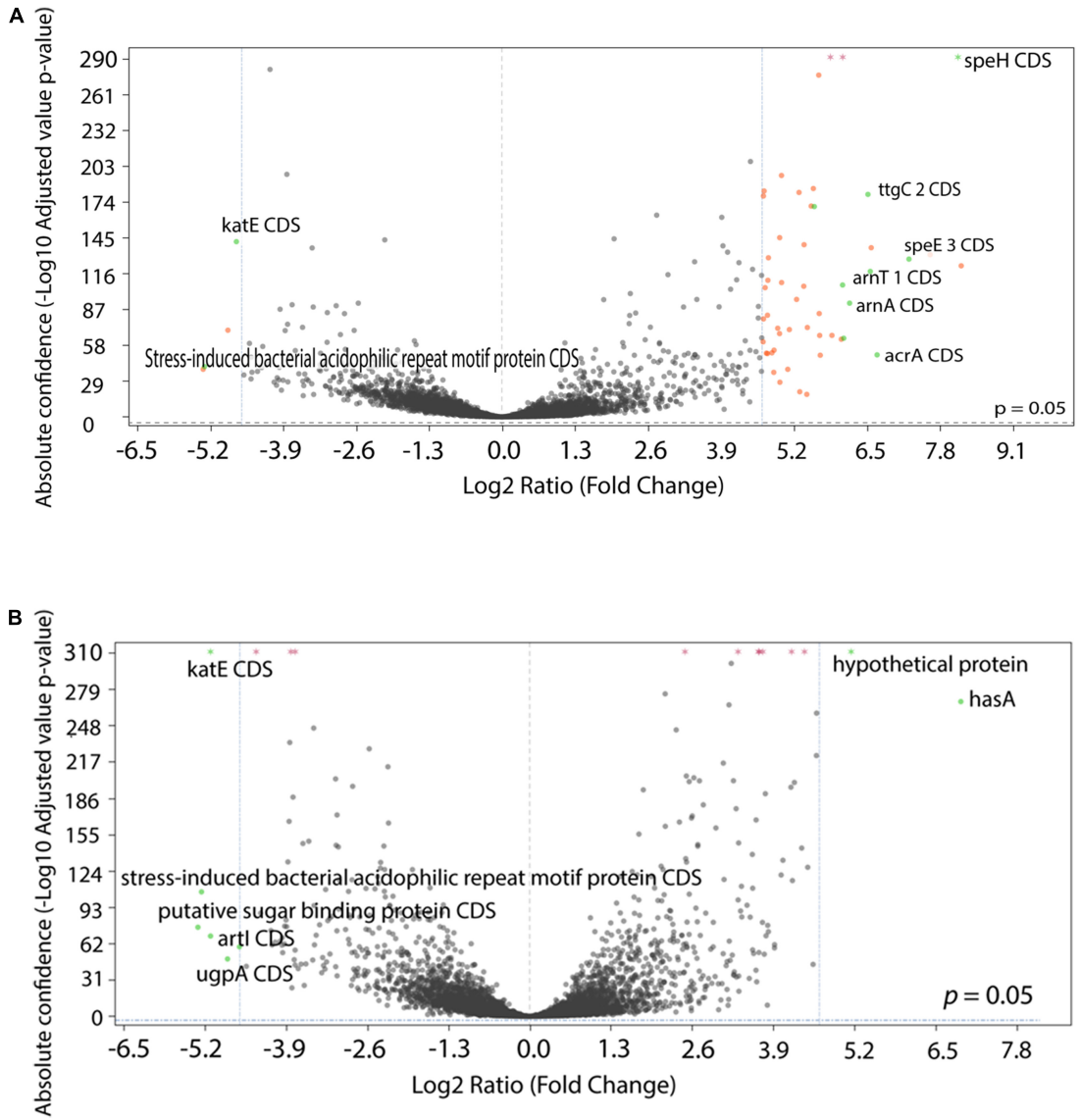


FIGURE 8 | Comparison of differentially expressed genes between conditions. **(A,B)** The statistical significance and magnitude of change for the differentially expressed genes are depicted as volcano plots. Day 21 isolates are shown here, cultivated two conditions: **(A)** comparison between colistin only and baseline, and **(B)** comparison between control isolates grown in LB medium and baseline. The log2 fold change is plotted against the significance (–Log10 adjusted *p* value). The horizontal dashed line represents a ratio of significance, set at 0.05. Each point represents a gene, with genes of interest are highlighted in green and labeled. Genes that are within the significant interest threshold but unlabelled are orange. Genes that do not meet the criteria are colored in black.

decreasing virulence for strains evolved in combination with colistin.

The downregulation of expression of the RNA polymerase-encoding gene *rpoS* in response to the drugs colistin and metronidazole is surprising, especially considering its role as a major stress-response regulator: it was expected that this gene would be upregulated in isolates cultured with antibiotics in the morbidostat. It is well documented that *rpoS* is expressed mainly during the stationary growth phase (Kojic and Venturi, 2001), and it may be that the continuous changes in growth kinetics in the morbidostat due to the addition of colistin at regular intervals may have led to an alternative transcriptomic profile. Alternatively, the reduced expression of *rpoS* in isolates exposed to colistin and metronidazole over 21 days, relative to the baseline strain PA77 might have been an adaptation mechanism to increase resistance to colistin. Indeed, previous studies have reported that *rpoS* mutants of *P. aeruginosa* produced more biofilm, and biofilms produced were much more resistant to being killed by tobramycin than wild-type *P. aeruginosa* biofilms (Whiteley et al., 2001). They also found enhanced flagellar motility exhibited by the *rpoS* mutants, which led to increase in biomass in biofilm, similar to what we have seen in our study. Increased and better developed biofilm protect against penetration of colistin (Ciofu and Tolker-Nielsen, 2019) with

extracellular polymeric substances reported to be the greatest barrier to diffusion for drug penetration into various bacteria (Wei and Ma, 2013). We see more biofilm produced in isolates exposed to colistin, either as a single treatment or in combination, than with metronidazole only. As deletion of this transcription factor has several important downstream effects (Lombardo et al., 2004), its reduced expression is likely to contribute to the altered transcriptional profile observed in the evolved isolates. *RpoS* controls close to 800 genes (Schuster et al., 2004) including several virulence factors such as pyocyanin, exotoxin A, LasA and LasB elastases, and exoenzyme S, and affects the expression of genes related to quorum sensing. Some of these genes are downregulated in our morbidostat-generated isolates, such as heat shock protein *dnaK*, and chaperonins *groES* and *groEL*, although not to a significant level (**Supplementary Table 2**).

The semi-automated morbidostat provides several advantages over traditional serial transfer evolution experiments. It enables us to simulate a clinical situation, as it represents compartments of infection within the human body that have not been fully eradicated by antibiotic treatment. These compartments of infection are exposed to sub-lethal concentrations of antibiotic, contributing to general tolerance. Isolates cultured in the morbidostat received sub-inhibitory doses of colistin, which creates a constant level of antibiotic pressure. The addition of metronidazole, which has no bactericidal effect on *P. aeruginosa*, but is a documented trigger for mutagenesis, leads to a significantly varied phenotype. In cases where colistin resistance develops, virulence potential decreases. However, isolates are able to survive and grow in human serum longer than their colistin-sensitive counterparts. Alongside the ability to maintain a higher number of viable cells in biofilm, the results of our work indicate the potential of a change in the course of infection due to significant evolutionary events under antibiotic treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: ebi.ac.uk; PRJEB41763.

AUTHOR CONTRIBUTIONS

MW and MJ designed the initial study setup. MJ, BJ, MW, SS, and SP selected strains, designed, and wrote protocols for the phenotypic experiments. MJ, BJ, MH, and VU completed phenotypic experiments, with MJ and MW analysing resulting data. MJ, MH, and VU performed wet lab work for NGS sequencing. MJ, MW, and AA performed NGS data analysis. US performed statistical analysis for experiments. MJ wrote the initial manuscript. MW compiled the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The work was supported by the German Center for Infection Research (grant number: TTU 08.702).

ACKNOWLEDGMENTS

We thank Libera Lo Presti, Michael Sonnabend, and Erwin Bohn for their contributions to the manuscript. We would also like to thank Richard Neher for his support with the morbidostat. We also extend our gratitude to our colleagues at AG Peschel for their advice and guidance regarding the experiments in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.619542/full#supplementary-material

REFERENCES

Ballok, A. E., and O'Toole, G. A. (2013). Pouring salt on a wound: *Pseudomonas aeruginosa* virulence factors alter Na+ and Cl– flux in the lung. *J. Bacteriol.* 195, 4013–4019. doi: 10.1128/JB.00339-13

Bolard, A., Schniederjans, M., Haüssler, S., Triponney, P., Valot, B., Plésiat, P., et al. (2019). Production of norspermidine contributes to aminoglycoside resistance in pmrAB mutants of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother* 63:e01044-19. doi: 10.1128/AAC.01044-19

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

Bravo, D., Silva, C., Carter, J. A., Hoare, A., Álvarez, S. A., Blondel, C. J., et al. (2008). Growth-phase regulation of lipopolysaccharide O-antigen chain length influences serum resistance in serovars of *Salmonella*. *J. Med. Microbiol.* 57, 938–946. doi: 10.1099/jmm.0.47848-0

Bricio-Moreno, L., Sheridan, V. H., Goodhead, I., Armstrong, S., Wong, J. K. L., Waters, E. M., et al. (2018). Evolutionary trade-offs associated with loss of PmrB function in host-adapted *Pseudomonas aeruginosa*. *Nat. Commun.* 9, 1–12. doi: 10.1038/s41467-018-04996-x

Burns, S. M., and Hull, S. I. (1998). Comparison of loss of serum resistance by defined lipopolysaccharide mutants and an acapsular mutant of uropathogenic *Escherichia coli* O75:K5. *Infect. Immun.* 66, 4244–4253.

Chou, H. T., Kwon, D.-H., Hegazy, M., and Lu, C.-D. (2008). Transcriptome analysis of agmatine and putrescine catabolism in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* 190, 1966–1975. doi: 10.1128/JB.01804-07

Ciofu, O., and Tolker-Nielsen, T. (2019). Tolerance and resistance of *Pseudomonas aeruginosa* biofilms to antimicrobial agents—How P. aeruginosa can escape antibiotics. *Front. Microbiol.* 10:913. doi: 10.3389/fmicb.2019.00913

Dößelmann, B., Willmann, M., Steglich, M., Bunk, B., Nübel, U., Peter, S., et al. (2017). Rapid and consistent evolution of colistin resistance in

extensively drug-resistant *Pseudomonas aeruginosa* during morbidostat culture. *Antimicrob. Agents Chemother.* 61:e00043-17. doi: 10.1128/AAC.00043-17

Esposito, E. P., Cervoni, M., Bernardo, M., Crivaro, V., Cuccurullo, S., Imperi, F., et al. (2018). Molecular epidemiology and virulence profiles of colistin-resistant *Klebsiella pneumoniae* Blood isolates from the hospital agency "Ospedale dei Colli," Naples, Italy. *Front. Microbiol.* 9:1463. doi: 10.3389/fmicb.2018.01463

EUCAST (2020). *Clinical Breakpoints and Dosing of Antibiotics*. Available online at: https://www.eucast.org/clinical_breakpoints/ accessed May 16, 2020).

European Centre for Disease Prevention and Control (2020). *Antimicrobial Resistance and Healthcare-Associated Infections - Annual Epidemiological Report 2014 [2012 Data]*. Available online at: https://www.ecdc.europa.eu/en/publications-data/antimicrobial-resistance-and-healthcare-associated-infections-annual (accessed June 19, 2020).

Falagas, M. E., Kasiakou, S. K., and Saravolatz, L. D. (2005). Colistin: the revival of polymyxins for the management of multidrug-resistant gram-negative bacterial infections. *Clin. Infect. Dis.* 40, 1333–1341. doi: 10.1086/429323

Ghadaksaz, A., Imani Fooladi, A. A., Hosseini, H. M., and Amin, M. (2015). The prevalence of some *Pseudomonas* virulence genes related to biofilm formation and alginate production among clinical isolates. *J. Appl. Biomed.* 13, 61–68. doi: 10.1016/j.jab.2014.05.002

Hocquet, D., and Bertrand, X. (2014). Metronidazole increases the emergence of ciprofloxacin- and amikacin-resistant *Pseudomonas aeruginosa* by inducing the SOS response. *J. Antimicrob. Chemother.* 69, 852–854. doi: 10.1093/jac/dkt435

Hong, M., and Payne, S. M. (1997). Effect of mutations in *Shigella* flexneri chromosomal and plasmid-encoded lipopolysaccharide genes on invasion and serum resistance. *Mol. Microbiol.* 24, 779–791. doi: 10.1046/j.1365-2958.1997.3731744.x

Hood, M. I., Becker, K. W., Roux, C. M., Dunman, P. M., and Skaar, E. P. (2013). Genetic determinants of intrinsic colistin tolerance in *Acinetobacter baumannii*. *Infect. Immun.* 81, 542–551. doi: 10.1128/IAI.00704-12

Jarrell', K. F., and Kropinski, A. M. (1982). The virulence of protease and cell surface mutants of *Pseudomonas aeruginosa* for the larvae of *Galleria mellonella*. *J. Invertebr. Pathol.* 39, 395–400.

Javed, M., Ueltzhoeffer, V., Heinrich, M., Siegrist, H. J., Wildermuth, R., Lorenz, F.-R., et al. (2018). Colistin susceptibility test evaluation of multiple-resistance-level *Pseudomonas aeruginosa* isolates generated in a morbidostat device. *J. Antimicrob. Chemother.* 73, 3368–3374. doi: 10.1093/jac/dky337

Johnson, L., Mulcahy, H., Kanevets, U., Shi, Y., and Lewenza, S. (2012). Surface-Localized spermidine protects the *Pseudomonas aeruginosa* outer membrane from antibiotic treatment and oxidative stress. *J. Bacteriol.* 194, 813–826. doi: 10.1128/JB.05230-11

Kojic, M., and Venturi, V. (2001). Regulation of rpoS gene expression in *Pseudomonas*: involvement of a TetR family regulator. *J. Bacteriol.* 183, 3712–3720. doi: 10.1128/JB.183.12.3712-3720.2001

Kunert, A., Losse, J., Gruszin, C., Hühn, M., Kaendler, K., Mikkat, S., et al. (2007). Immune evasion of the human pathogen *Pseudomonas aeruginosa*: elongation factor Tuf Is a factor H and plasminogen binding protein. *J. Immunol.* 179, 2979–2988. doi: 10.4049/jimmunol.179.5.2979

Land, K. M., and Johnson, P. J. (1999). Molecular basis of metronidazole resistance in pathogenic bacteria and protozoa. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* 2, 289–294. doi: 10.1054/drup.1999.0104

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

Lee, J.-S., Heo, Y.-J., Lee, J. K., and Cho, Y.-H. (2005). KatA, the major catalase, is critical for osmoprotection and virulence in *Pseudomonas aeruginosa* PA14. *Infect. Immun.* 73, 4399–4403. doi: 10.1128/IAI.73.7.4399-4403.2005

Lee, J.-Y., Park, Y. K., Chung, E. S., Na, I. Y., and Ko, K. S. (2016). Evolved resistance to colistin and its loss due to genetic reversion in *Pseudomonas aeruginosa*. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep25543

Lo Sciuto, A., Cervoni, M., Stefanelli, R., Mancone, C., and Imperi, F. (2020). Effect of lipid A aminoarabinosylation on *Pseudomonas aeruginosa* colistin resistance and fitness. *Int. J. Antimicrob. Agents* 55:105957. doi: 10.1016/j.ijantimicag.2020.105957

Löfmark, S., Edlund, C., and Nord, C. E. (2010). Metronidazole is still the drug of choice for treatment of anaerobic infections. *Clin. Infect. Dis.* 50, S16–S23. doi: 10.1086/647939

Lombardo, M.-J., Aponyi, I., and Rosenberg, S. M. (2004). General stress response regulator RpoS in adaptive mutation and amplification in *Escherichia coli*. *Genetics* 166, 669–680. doi: 10.1534/genetics.166.2.669

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Magiorakos, A.-P., Srinivasan, A., Carey, R. B., Carmeli, Y., Falagas, M. E., Giske, C. G., et al. (2012). Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* 18, 268–281. doi: 10.1111/j.1469-0691.2011.03570.x

Maurice, N. M., Bedi, B., and Sadikot, R. T. (2018). *Pseudomonas aeruginosa* biofilms: host response and clinical implications in lung infections. *Am. J. Respir. Cell Mol. Biol.* 58, 428–439. doi: 10.1165/rcmb.2017-0321TR

McPhee, J. B., Lewenza, S., and Hancock, R. E. W. (2003). Cationic antimicrobial peptides activate a two-component regulatory system, PmrA-PmrB, that regulates resistance to polymyxin B and cationic antimicrobial peptides in *Pseudomonas aeruginosa*. *Mol. Microbiol.* 50, 205–217. doi: 10.1046/j.1365-2958.2003.03673.x

Miajlovic, H., and Smith, S. G. (2014). Bacterial self-defence: how *Escherichia coli* evades serum killing. *FEMS Microbiol. Lett.* 354, 1–9. doi: 10.1111/1574-6968.12419

Mulcahy, L. R., Isabella, V. M., and Lewis, K. (2014). *Pseudomonas aeruginosa* biofilms in disease. *Microb. Ecol.* 68, 1–12. doi: 10.1007/s00248-013-0297-x

Necchi, F., Saul, A., and Rondini, S. (2017). Development of a high-throughput method to evaluate serum bactericidal activity using bacterial ATP measurement as survival readout. *PLoS One* 12:e0172163. doi: 10.1371/journal.pone.0172163

Nowicki, E. M., O'Brien, J. P., Brodbelt, J. S., and Trent, M. S. (2014). Characterization of *Pseudomonas aeruginosa* LpxT reveals dual positional lipid A kinase activity and coordinated control of outer membrane modification. *Mol. Microbiol.* 94, 728–741. doi: 10.1111/mmi.12796

Ohno, A., Isii, Y., Tateda, K., Matumoto, T., Miyazaki, S., Yokota, S., et al. (1995). Role of LPS length in clearance rate of bacteria from the bloodstream in mice. *Microbiology* 141, 2749–2756. doi: 10.1099/13500872-141-10-2749

Qu, L., She, P., Wang, Y., Liu, F., Zhang, D., Chen, L., et al. (2016). Effects of norspermidine on *Pseudomonas aeruginosa* biofilm formation and eradication. *MicrobiologyOpen* 5, 402–412. doi: 10.1002/mbo3.338

Rice, L. B. (2008). Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No ESKAPE. *J. Infect. Dis.* 197, 1079–1081. doi: 10.1086/533452

Romano, K. P., Warrier, T., Poulsen, B. E., Nguyen, P. H., Loftis, A. R., Saebi, A., et al. (2019). Mutations in pmrB confer cross-resistance between the LptD inhibitor POL7080 and colistin in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* 63:e00511-19. doi: 10.1128/AAC.00511-19

Schuster, M., Hawkins, A. C., Harwood, C. S., and Greenberg, E. P. (2004). The *Pseudomonas aeruginosa* RpoS regulon and its relationship to quorum sensing. *Mol. Microbiol.* 51, 973–985. doi: 10.1046/j.1365-2958.2003.03886.x

Short, F. L., Sario, G. D., Reichmann, N. T., Kleanthous, C., Parkhill, J., and Taylor, P. W. (2020). Genomic profiling reveals distinct routes to complement resistance in *Klebsiella pneumoniae*. *Infect. Immun.* 88:e00043-20. doi: 10.1128/IAI.00043-20

Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L., and Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.* 44, 101–105. doi: 10.1038/ng.1034

Tsai, C. J.-Y., Loh, J. M. S., and Proft, T. (2016). Galleria mellonella infection models for the study of bacterial diseases and for antimicrobial drug testing. *Virulence* 7, 214–229. doi: 10.1080/21505594.2015.1135289

Wassermann, T., Meinike Jørgensen, K., Ivanyshyn, K., Bjarnsholt, T., Khademi, S. M. H., Jelsbak, L., et al. (2016). The phenotypic evolution of *Pseudomonas aeruginosa* populations changes in the presence of subinhibitory concentrations of ciprofloxacin. *Microbiology* 162, 865–875. doi: 10.1099/mic.0.000273

Wei, Q., and Ma, L. Z. (2013). Biofilm matrix and its regulation in *Pseudomonas aeruginosa*. *Int. J. Mol. Sci.* 14, 20983–21005. doi: 10.3390/ijms141020983

Whiteley, M., Bangera, M. G., Bumgarner, R. E., Parsek, M. R., Teitzel, G. M., Lory, S., et al. (2001). Gene expression in *Pseudomonas aeruginosa* biofilms. *Nature* 413, 860–864. doi: 10.1038/35101627

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* 3:e000132. doi: 10.1099/mgen.0.000132

Willmann, M., Klimek, A. M., Vogel, W., Liese, J., Marschal, M., Autenrieth, I. B., et al. (2014). Clinical and treatment-related risk factors for nosocomial colonisation with extensively drug-resistant *Pseudomonas aeruginosa* in a haematological patient population: a matched case control study. *BMC Infect. Dis.* 14:650. doi: 10.1186/s12879-014-0650-9

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.

Copyright © 2021 Javed, Jentzsch, Heinrich, Ueltzhoeffer, Peter, Schoppmeier, Angelov, Schwarz and Willmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

REVIEW
published: 04 February 2021
doi: 10.3389/fmicb.2021.612675

Evolution of Drug-Resistant *Mycobacterium tuberculosis* Strains and Their Adaptation to the Human Lung Environment

Anna Allué-Guardia, Juan I. García and Jordi B. Torrelles**

Population Health Program, Tuberculosis Group, Texas Biomedical Research Institute, San Antonio, TX, United States

In the last two decades, multi (MDR), extensively (XDR), extremely (XXDR) and total (TDR) drug-resistant *Mycobacterium tuberculosis* (*M.tb*) strains have emerged as a threat to public health worldwide, stressing the need to develop new tuberculosis (TB) prevention and treatment strategies. It is estimated that in the next 35 years, drug-resistant TB will kill around 75 million people and cost the global economy $16.7 trillion. Indeed, the COVID-19 pandemic alone may contribute with the development of 6.3 million new TB cases due to lack of resources and enforced confinement in TB endemic areas. Evolution of drug-resistant *M.tb* depends on numerous factors, such as bacterial fitness, strain's genetic background and its capacity to adapt to the surrounding environment, as well as host-specific and environmental factors. Whole-genome transcriptomics and genome-wide association studies in recent years have shed some insights into the complexity of *M.tb* drug resistance and have provided a better understanding of its underlying molecular mechanisms. In this review, we will discuss *M.tb* phenotypic and genotypic changes driving resistance, including changes in cell envelope components, as well as recently described intrinsic and extrinsic factors promoting resistance emergence and transmission. We will further explore how drug-resistant *M.tb* adapts differently than drug-susceptible strains to the lung environment at the cellular level, modulating *M.tb*–host interactions and disease outcome, and novel next generation sequencing (NGS) strategies to study drug-resistant TB.

Keywords: *Mycobacterium tuberculosis*, drug resistance, evolution, bacterial–host interactions, next generation sequencing

OPEN ACCESS

Edited by:
Jianying Gu,
College of Staten Island,
United States

Reviewed by:
John T. Belisle,
Colorado State University,
United States
Oana Dumitrescu,
Hospices Civils de Lyon, France

****Correspondence:***
Anna Allué-Guardia
aallueguardia@txbiomed.org
Jordi B. Torrelles
jtorrelles@txbiomed.org

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

Received: *30 September 2020*
Accepted: *15 January 2021*
Published: *04 February 2021*

Citation:
Allué-Guardia A, García JI and Torrelles JB (2021) Evolution of Drug-Resistant Mycobacterium tuberculosis Strains and Their Adaptation to the Human Lung Environment. Front. Microbiol. 12:612675. doi: 10.3389/fmicb.2021.612675

INTRODUCTION

Tuberculosis (TB) kills one person every 21s, with ∼10 million cases and ∼1.5 million attributed deaths in 2018 (WHO, 2019). Caused by airborne pathogen *Mycobacterium tuberculosis* (*M.tb*), TB is the top disease killer worldwide due to a single infectious agent. After inhalation, *M.tb* reaches the alveolar space and is bathed in alveolar lining fluid (ALF), being in intimate contact with soluble components of the lung mucosa before interacting with the cellular compartment, including alveolar macrophages (AM) and other immune cells (Arcos et al., 2011; Torrelles and Schlesinger, 2017). This *M.tb*–host interplay during the initial stages of infection will determine TB disease outcome, driving clearance or progression to active (pulmonary or extrapulmonary

TB) or latent TB, although the exact bacterial/host determinants and mechanisms that promote a successful infection are still poorly understood. The success of *M.tb* in establishing infection and evading the host immune system is in part due to its unique and dynamic cell envelope, composed mainly of lipids and carbohydrates, which changes to adapt to the different lung microenvironments in response to local environmental cues, and also protects the pathogen against harsh environments and many antimicrobial drugs due to its low permeability (Garcia-Vilanova et al., 2019; Maitra et al., 2019). And thus, most combined treatments for active TB work disrupting the mycobacterial cell envelope.

Standard treatment for drug-susceptible TB consists of administering four first-line drugs for 6 months (2 months of isoniazid, rifampicin, ethambutol and pyrazinamide, followed by 4 months of isoniazid and rifampicin), with a success rate of 85% (WHO, 2019). However, the worldwide emergence of drug-resistant strains in recent years, especially in TB endemic areas (Ellner, 2008; Lange et al., 2019), poses a global threat and compromises the goal of the End TB Strategy of reducing 90% the TB cases by 2035. Different causes have led to increased drug resistance, including insufficient healthcare infrastructure, prescription of the wrong treatment (either wrong dose or treatment length), poor quality drugs or drug unavailability, poor adherence to therapy, or *M.tb* re-infection, among others (Zaman, 2010). Indeed, in 2018, approximately half a million people developed drug-resistant TB (DR-TB) and of these, 78% were multidrug-resistant (MDR-TB), and only one in three confirmed cases were enrolled in treatment (WHO, 2019). Different categories of drug-resistant *M.tb* strains have been defined by the WHO: multi (MDR), extensively (XDR), and extremely (XXDR) drug-resistant TB. MDR-TB, which represent a 3.4% of new TB cases worldwide, are resistant to at least the two first-line drugs isoniazid and rifampicin, and require 9–20 months of costly treatment using second-line drugs, with decreased success rate of 56% when compared to susceptible TB (WHO, 2019). XDR-TB are also resistant to rifampicin and isoniazid, plus fluoroquinolone and at least one of the three injectable second-line drugs (amikacin, kanamycin, or capreomycin), with even a lower treatment success rate of 39% (WHO, 2019). Lastly, XXDR strains are resistant to all first and second-line drugs. More recently, a new category of *M.tb* strains not yet recognized by the WHO have been identified in Italy, India and Iran (Velayati et al., 2009b; Loewenberg, 2012), and named totally drug-resistant TB (TDR-TB) for being resistant to all tested antibiotics plus some of the ones currently in the discovery pipeline (Loewenberg, 2012; Parida et al., 2015). The emergence of these potentially incurable strains stresses the urgent need to develop new drug regimens and/or alternative anti-TB strategies to combat these superbugs (Lange et al., 2018).

Only a few studies have explored the cell envelope composition of drug-resistant *M.tb* strains and their metabolic responses and adaptation to the different environmental pressures during the course of pulmonary infection, which makes it challenging to develop new and effective treatments (Velayati et al., 2009b). Due to an altered cell envelope composition of drug-resistant isolates (Velayati et al., 2013), one can hypothesize that drug-susceptible and drug-resistant strains will adapt differently to the lung microenvironments, driving different infection outcomes. In recent years, new developments in next generation sequencing (NGS) technologies have allowed investigators to gain further insight not only in specific genetic determinants of drug resistance but have also shed some light on how particular *M.tb* strains adapt and interact with the host during infection through genome-wide transcriptomics, which is key in finding novel bacterial determinants that can be targeted in the development of new anti-TB drugs or host-directed therapies (Jeanes and O'Grady, 2016).

In this review, we will discuss different factors driving the emergence of drug-resistant *M.tb* strains, as well as what is known about the adaptation of MDR strains to the different lung environments encountered during the course of infection, and describe recent NGS strategies that could be used to gain a better understanding on *M.tb*–host interactions and key determinants involved in MDR evolution.

ORIGIN AND EVOLUTION OF DRUG RESISTANCE

Mycobacterium tuberculosis is a highly specialized human intracellular pathogen with an extremely conserved genome driving a paradigm of a near-perfect host-pathogen relationship (Brites and Gagneux, 2015). Currently, phylogenetic inferences reveal that there are seven global lineages of *M.tb* strains that have co-evolved with human populations under sympatric and allopatric host-pathogen combinations (Gagneux et al., 2006a; Al-Saeedi and Al-Hajoj, 2017). These *M.tb* strain lineages differ in their geographic distribution, biological fitness, virulence and their propensity to acquire drug resistance; specifically, lineages 2 to 4 have been associated with greater disease burden and drug resistance compared to ancient lineages, e.g., lineages 1, 5, and 6 (Al-Saeedi and Al-Hajoj, 2017; Nguyen et al., 2018).

The emergence of drug resistance in *M.tb* can be divided in two main factors and their interactions: extrinsic factors, which are related to social determinants of TB in populations and the quality of TB control and prevention services; and intrinsic factors, accounting for those related to the acquisition of genetic mutations in drug resistance-associated genes (Gygli et al., 2017; Nguyen et al., 2018; Swain et al., 2020). Regarding the intrinsic factors, mutations in genes coding for drug targets or drug activating enzymes are the primary mode of drug resistance, and they arise mainly through single nucleotide polymorphisms (SNPs) and insertion-deletions (indels). Unlike other bacterial pathogens, acquisition of drug resistance through horizontal gene transfer is not consistently reported in *M.tb* (Namouchi et al., 2012; Dookie et al., 2018). In addition, drug resistance in *M.tb* is commonly believed to be caused by single-step chromosomal mutations. However, there are now evidences suggesting that, at least for certain anti-TB drugs, acquired drug resistance is the result of a stepwise acquisition and fixation of mutations leading to a gradual increase in resistance, initiated with the acquisition of isoniazid resistance, subsequently followed by rifampicin or ethambutol resistance (Cohen et al., 2015). Indeed,

resistance to isoniazid through the *KatG*S315T mutation is a common pattern that precedes rifampicin resistance and is conserved globally regardless of the *M.tb* lineage, geographic region and/or time (Manson et al., 2017). DR-TB can occur as primary drug resistance, when a person is directly infected by a drug-resistant *M.tb* strain, or as secondary or acquired drug resistance, occurring due to the acquisition of resistance-conferring mutations during failed treatment of drug-susceptible TB (Dookie et al., 2018). The latter is likely associated to *M.tb* metabolic adaptation to the host lung environment.

In addition to resistance caused by target mutations, several distinctive mechanisms of anti-*M.tb* innate resistance have been also described (Singh et al., 2020). These include: drug accessibility to the target due to the low permeability of the *M.tb* cell envelope, modification of drugs by *M.tb* enzymes, existence of *M.tb* efflux pumps removing out drugs that were able to cross the *M.tb* cell envelope, *M.tb* modulation of its gene expression to adapt to the drug's effects (or to its presence), and a phenotypic drug tolerance linked to a state of slow growth rate and metabolic shut down. In this regard, during *M.tb* infection certain bacterial subpopulations, known as persisters, can become phenotypically tolerant to antimycobacterial drugs without acquiring genetic mutations (Kester and Fortune, 2014). This reversible phenomenon is usually induced by external stresses such as hypoxia or drug treatment, among others factors (Vilcheze and Jacobs, 2019), and is associated with a nonreplicating status of the bacilli. Several *M.tb* factors and metabolic traits are linked to *M.tb* persistence (Keren et al., 2011; Lee et al., 2019), although the evidence suggests that persistence consists of an array of heterogenous physiological states and mechanisms (Hartman et al., 2017). Mechanisms of *M.tb* persistence are still being elucidated; however, different bacterial metabolic status during early stages of the infection might explain why some *M.tb* subpopulations can become phenotypically drug-resistant.

Indeed, a novel genetically encoded mechanism that causes reversible drug tolerance was recently described in *M.tb* (Bellerose et al., 2019; Safi et al., 2019; Vargas and Farhat, 2020). This phase variation phenomenon is caused by transient frameshift mutations in a 5' homopolymeric region in the *glpK* gene, which encodes the glycerol-3-kynase required for *M.tb* glycerol catabolism. This phenomenon has been observed in several clinical isolates which, as a consequence, generate bacterial variants (small smooth colonies) that exhibit a drug-tolerant phenotype easily reversible through the introduction of additional insertions/deletions in the same *glpK* region. This frameshift mutations are particularly enriched in MDR- and XDR-*M.tb* strains, and suggest that variation in GlpK might contribute to the development of drug resistance. Thus, limited efficacy of current TB treatments suggests that heterogeneity of both host and mycobacterial physiologies in the different lung compartments during the different stages of the infection can influence the emergence of persisters and ultimately, increased drug resistance (Warner and Mizrahi, 2006; Ben-Kahla and Al-Hajoj, 2016; Boldrin et al., 2020). Novel antimycobacterial drugs that target persisters populations are critical in order to shorten TB treatment and decrease the emergence and transmission of MDR-*M.tb* strains (Zhang et al., 2012; Mandal et al., 2019). Conversely, mixed clonal infections and/or genetic heterogeneity of *M.tb* populations with different drug susceptibility profiles within a patient can also lead to disparate responses during treatment, promoting drug resistance (Liu et al., 2015).

Factors Contributing to the Emergence of Drug Resistance

The mechanisms and pathways that result in the emergence and subsequent fixation of *M.tb* resistant strains and its dynamics are not fully understood. However, evidence suggest an interplay of several mechanisms involved during drug selection pressure, including clonal interference, mutation rates, efflux pumps, compensatory mutations, and epistasis (Al-Saeedi and Al-Hajoj, 2017). Exposure of *M.tb* to drugs induces a bacterial stress response while exerting a drug selection pressure; thus, only those *M.tb* strains able to adapt will prevail, initiating a competitive selection process between *M.tb* clones that may acquire different beneficial mutations to survive (clonal interference). In this regard, *M.tb* strain lineages can present different mutation rates and different capacity to acquire drug resistance (Ford et al., 2013), where clones of the Beijing/East Asian lineage (lineage 2) are associated to an hypermutability phenotype mediated through polymorphisms in anti-mutator genes (*mut)* involved in DNA repair systems (Ebrahimi-Rad et al., 2003). Recently, DNA methylation has been proposed as a mechanism for phenotypic plasticity in *M.tb*, aiding rapid adaptation to changing environmental pressures, with heterogeneous DNA methylation reported in MTBC clinical isolates (Modlin et al., 2020), which could have potential implications in the development of drug resistance. Further, because of the drug-induced stress response of *M.tb*, *M.tb* efflux pumps are rapidly upregulated within hours after drug exposure, being responsible of conferring a low-level resistance profile (described in detail below). Interestingly, efflux pump inhibitors reduce the minimum inhibitory concentrations (MICs) of key drugs such as isoniazid, rifampicin, linezolid and some fluoroquinolones (Escribano et al., 2007; Machado et al., 2012). Thus, this efflux pump mediated low-level resistance confers a selective advantage and allows *M.tb* to survive and replicate under sub-optimal drug concentrations, until further development of classical resistance-associated mutations that confer clinical drug-resistant phenotypes (Fonseca et al., 2015; Dheda et al., 2017).

Although beneficial against anti-TB treatment, drug resistance mutations may also impose a fitness cost to *M.tb* survival, as they can compromise genes involved in essential mycobacterial cell functions. This subject is a matter of debate aiming to elucidate the mechanisms involved in impaired fitness, transmissibility and virulence of DR-*M.tb* strains. The concept of reduced virulence of MDR-*M.tb* strains comes from early experiments in the Guinea pig infection model (Middlebrook and Cohn, 1953). More recently, clinical isolates with common mutations for isoniazid, rifampicin, and streptomycin were associated with low or absent *in vivo* fitness cost (Comas et al., 2011). Thus, clinical isolates undergo mutations to overcome fitness deficits associated to drug

resistance, setting the basis for the compensatory evolutionary role in the spread of DR-TB (Nguyen et al., 2018). There is a wide diversity of resistance-conferring mutations in *M.tb*; however, during selective pressure, specific mutations in clinical isolates are associated with a high-level of resistance without losing fitness or transmissibility (Casali et al., 2014). Different compensatory mutations were demonstrated for *rpoA* and *rpoC* in rifampicin-resistant *rpoB* mutants (Comas et al., 2011; de Vos et al., 2013), for *ahpC* in isoniazid-resistant *KatG* mutants (Sherman et al., 1996), and 16S RNA for aminoglycoside resistance (Shcherbakov et al., 2010). This evidence supports the successful expansion of MDR and XDR-*M.tb* strains worldwide (Merker et al., 2018; Cohen et al., 2019a).

Epistasis, a form of interaction between genes or mutations that influences a phenotype, is also thought to drive the evolution of DR-TB (Borrell et al., 2013; Koch et al., 2014). In fact, some *M.tb* compensatory mutations are associated with epistatic interactions. Epistasis is conceptually linked to fitness and is divided into negative and positive epistasis; the latter meaning that the interaction between genes or mutations has a smaller fitness cost compared to the genetic determinants alone. Different positive and negative epistasis exist in *M.tb* between drug-resistant mutations and between drug resistance-associated mutations and compensatory mutations (Phillips, 2008; Nguyen et al., 2018). In this context, different lineages of *M.tb* carrying identical rifampicin-associated mutations showed different levels of fitness cost, supporting the idea that epistasis is influenced by *M.tb* strain's genetic background (Gagneux et al., 2006b; Borrell and Gagneux, 2011; Li et al., 2017). Epistatic interactions may also determine the order in which drug resistance mutations and compensatory mutations arise, playing a key role in defining the evolutionary processes toward acquisition of MDR-TB (Salverda et al., 2011; Muller et al., 2013).

ADAPTATION OF DRUG-RESISTANT *M.tb* STRAINS TO THE HUMAN LUNG ENVIRONMENT

Mycobacterium tuberculosis evolved from a mycobacterial ancestor to effectively infect and persist in the human host (Brites and Gagneux, 2015; Jia et al., 2017). Its higher hydrophobicity compared to other mycobacterial species might have contributed to increased transmission through aerosolization (Jankute et al., 2017). After being inhaled and during the course of infection, *M.tb* gets in contact with very different host lung microenvironments, which can be extracellular (during the initial stages of infection and before interacting with alveolar phagocytes and other immune cells, after escaping necrotic cells, or after re-activation in cavities); or intracellular (in alveolar phagocytes such as alveolar macrophages (AMs) during primary infection, or within AMs in granulomas during the latency stage) (**Figure 1**). In these different environments, *M.tb* has evolved to use host resources to its advantage and adapt its metabolism in order to evade the host immune system, survive, and establish a successful active or latent infection.

The *M.tb* cell envelope provides structural support and protection to osmotic changes, and is composed of four main layers: an inner plasma membrane and periplasmic space, a cell envelope core composed of peptidoglycan (PG) covalently linked to arabinogalactan (AG) and mycolic acids (MAs), the peripheral lipid layer [formed by non-covalently linked lipids and glycolipids such as trehalose dimycolate (TDM), phthiocerol dimycocerosates (PDIMs), mannose-capped lipoarabinomannan (ManLAM), sulfolipids (SLs), phosphatidyl-*myo*-inositol mannosides (PIMs), and phenolic glycolipids (PGLs), among others], and the outermost layer (also named capsule) (Kalscheuer et al., 2019). Although lipids and carbohydrates constitute ∼80% of the cell wall, proteins are also important components of the *M.tb* envelope, being studied as potential drug targets (Gu et al., 2003; Mawuenyega et al., 2005; Niederweis et al., 2010; Chiaradia et al., 2017; Hermann et al., 2020). Several described proteins regulate the permeability of the *M.tb* cell envelope, with important implications in drug resistance. Among many, we find mycobacterial membrane protein Large 3 (MmpL3), Rv3143/Rv1524 axis, and decaprenylphosphoryl-*D*-ribose 2′-epimerase (DprE), which are involved in the export and synthesis of the *M.tb* cell wall, regulating its permeability (Dong et al., 2020). Other *M.tb* cell wall proteins act directly inactivating the drug. This is the case of Rv2170, a putative acetyltransferase that acetylates isoniazid (INH), inducing its breakdown into acetylhydrazine and isonicotinic acid, thus defining a mechanism by which some *M.tb* strains could bypass INH toxicity (Arun et al., 2020).

Besides providing structural support, the *M.tb* cell envelope also plays a critical immunomodulatory role in the bacterium-host crosstalk, where several cell envelope outer molecules are known to participate in different stages of the infection with key implications in *M.tb* pathogenesis and development of drug resistance (Sonawane et al., 2012; Gago et al., 2018; Garcia-Vilanova et al., 2019; Kalscheuer et al., 2019). Indeed, the mycobacterial cell envelope is considered a highly dynamic structure, tightly regulated and remodeled in response to changing host environmental pressures. Consequently, a wide range of cell envelope compositions in different bacterial cells can be found during the course of infection (Kalscheuer et al., 2019). These will determine how *M.tb* interacts with the host immune system and its potential implications for antibiotic resistance (Dulberger et al., 2020).

However, little is known about the temporal changes of the *M.tb* cell envelope in response to local environmental cues and how these can lead to different disease outcomes, especially in DR-*M.tb* strains, since most of the data come from *in vitro* studies and/or specific stages of infection, not necessarily reflecting the actual temporal dynamics of this complex and heterogeneous structure. *M.tb*–host interactions and subsequent infection outcome might also be influenced by *M.tb* metabolic and/or physiologic status before and/or during infection. Below we will discuss recent developments on MDR-*M.tb* adaptation to the different human lung microenvironments, and how the dynamics of the *M.tb* cell envelope might determine disease outcome during the *M.tb*-host interplay.

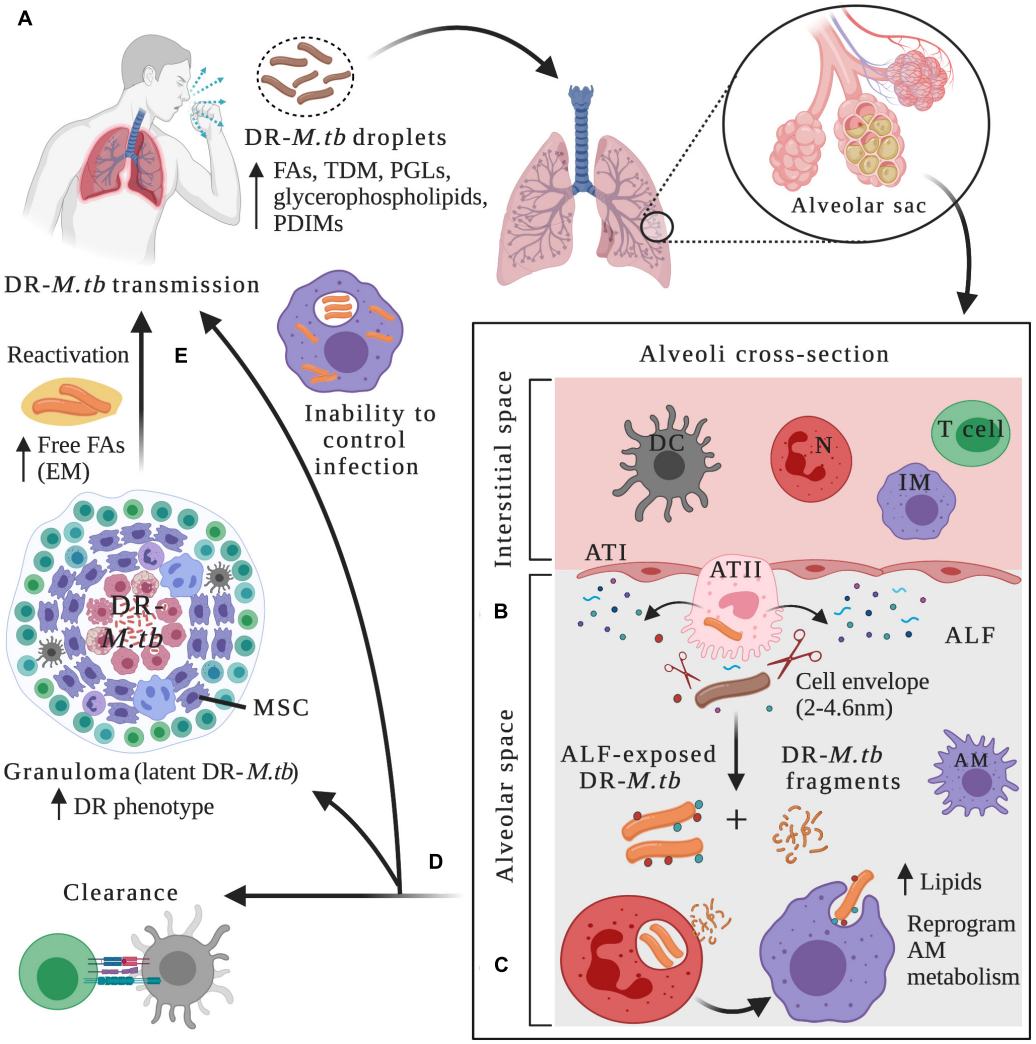


FIGURE 1 | Drug-resistant *M.tb*–host interactions within lung microenvironments at different stages of the infection. **(A)** After being in close contact with an individual with active TB disease, infection can be initiated through inhalation of drug-resistant (DR) *M.tb*-containing droplets. DR-*M.tb* bacilli contain altered levels of cell envelope lipids such as free fatty acids (FAs), trehalose dimycolate (TDM), Phthiocerol dimycocerosates (PDIMs), phenolic glycolipids (PGLs) and glycerophospholipids, among others. Upon bypassing upper respiratory tract barriers, DR-*M.tb* will ultimately reach the alveoli, a sac-like structure composed of a thin layer of alveolar epithelial cells type I (ATIs, with structural and gas exchange function) and alveolar epithelial cells type II (ATII, with secretor function) surrounded by capillaries. Alveolar macrophages (AMs) are resident phagocytes that populate the alveolar space, while the interstitial space surrounding the alveoli contains interstitial macrophages (IMs), dendritic cells (DCs), neutrophils (N), and T cells, among other host cells. **(B)** In the alveolar space, DR-*M.tb* bacilli first interact with host soluble innate components present in the alveolar lining fluid (ALF), where hydrolases (represented as scissors) can cleave and modify the *M.tb* cell envelope, releasing cell envelope fragments into the alveolar space. **(C)** Subsequently, ALF-modified DR-*M.tb* bacilli will interact with AMs (professional phagocytes) and/or with ATs (non-professional phagocytes), as well as with other host innate immune cells (e.g., neutrophils, DCs). Released DR-*M.tb* fragments are immunogenic and could attract neutrophils to the infection site driving local oxidative stress and inflammation, which could assist resident resting AMs to clear the infection. **(D)** The outcome of these initial interactions will resolve in ALF-exposed DR-*M.tb* clearance, establishment of a successful infection driving primary active TB disease, or a latent *M.tb* infection defined by *M.tb* persisters within granulomas, a niche that provide a protective environment against anti-TB drugs, thus increasing the DR-*M.tb* phenotype. Surrounding mesenchymal stem cells (MSCs) can also dampen immune responses and provide a protective intracellular environment for *M.tb* persistence. **(E)** Reactivation and subsequent progression to active TB disease can happen when granulomas fail to contain DR-*M.tb*, with extracellular DR-*M.tb* growth that leads to lung tissue destruction and cavity formation. It has been suggested that in this scenario, DR-*M.tb* secretes free fatty acids, creating some kind of extracellular matrix (EM) that further shields the DR-*M.tb* against TB drugs. Figure created with BioRender.com.

Initial *M.tb*–Host Interactions Within the Lung Mucosa

Mycobacterium tuberculosis infection requires close contact with an infectious individual with active TB, occurring through airborne transmission of droplets containing viable bacteria (**Figure 1A**). Infection in the lungs is initiated when the pathogen is inhaled and bypasses bacterial clearance mechanisms present in the first interaction points that act as barriers, including nose and sinuses, tracheobronchial tree and bronchioles, ultimately reaching the alveoli (Torrelles and Schlesinger, 2017)

(**Figure 1A**). In the alveolar space, *M.tb* first comes in close contact with soluble elements present in the lung alveolar mucosa, which is composed of surfactant lipids and a hypophase layer or ALF (**Figure 1B**). ALF contains surfactant proteins, hydrolytic enzymes, and complement proteins, among others, all critical components of the host soluble innate immune response. Indeed, ALF hydrolases are known to interact and modify the cell envelope of *M.tb*, significantly reducing the amount of two major virulence factors, ManLAM and TDM on the bacterial cell envelope surface (Arcos et al., 2011) (**Figure 1B** and **Table 1**). This ALF-driven changes on the *M.tb* cell envelope are correlated with a significantly better control of *M.tb* intracellular growth by professional phagocytes *in vitro* by increasing phagosome-lysosome fusion, as well as long-term effects *in vivo* (Arcos et al., 2011, 2015; Moliva et al., 2019) (**Table 1**), suggesting a critical role of these first interactions between *M.tb* and host innate soluble proteins in determining infection and disease outcome. In fact, *M.tb* exposure to ALF, besides modifying the *M.tb* cell envelope, can also drive differential *M.tb* infection outcome in alveolar epithelial cells (ATs). In this regard, two distinct subsets of human ALFs are defined, the ones that upon exposure drove *M.tb* to replicate faster (H-ALF), and the ones that upon exposure slowed down *M.tb* replication (L-ALF) within ATs. This *M.tb* differential replication within ATs is further correlated with an altered human ALF composition and function, where people with H-ALF could be more susceptible to *M.tb* infection (Scordo et al., 2019) (**Table 1**).

Indeed, altered ALF composition and functionality in certain human populations such as the elder, HIV-infected, diabetic and/or individuals with underlying chronic lung diseases might be a contributing factor related to increased susceptibility to *M.tb* infection (Moliva et al., 2014, 2019; Bell and Noursadeghi, 2018; Segura-Cerda et al., 2019), demonstrating the protective and unique role of ALF in determining *M.tb* infection control and/or progression to disease. In this context, the elderly population is characterized by a pro-inflammatory lung environment, with increased protein oxidation levels in the ALF linked to decreased soluble innate immune function (Moliva et al., 2014, 2019).

TABLE 1 | DS-*M.tb* and DR-*M.tb* interactions with the host at different stages of the infection.

Lung environments	DS-*M.tb*	DR-*M.tb*
Initial *M.tb*–host interactions ALF (e.g., hydrolases, surfactant proteins, etc.)	• Decrease of ManLAM and TDM in the cell envelope surface: increased P-L fusion in phagocytes (better control of *M.tb* intracellular growth). • Differential *M.tb* infection outcome in ATs after ALF exposure: H-ALF vs. L-ALF (based on ALF composition and functionality: host component). • Increased neutrophil's capacity to eliminate ALF-exposed *M.tb* (increased P-L fusion and higher levels of TNF and IL-8, but reduced oxidative burst).	• Thicker cell envelope. • Presumably altered levels of fatty acids, TDM, TMM, glycerophospholipids, GLs, PDIM, PGL. • Presumably altered protein composition. • Presumably altered ALF-driven cell envelope modifications.
Initial stages of infection Macrophages, Neutrophils, ATs	• PAMPs in the *M.tb* cell envelope recognized by macrophages, DCs and other immune cells through different receptors. • Mannose-containing molecules in the cell envelope promote recognition and survival inside phagocytic cells. Host defense and immunomodulatory roles. • Efflux pumps correlated with development of drug resistance. • ManLAM, TDM, and PDIM levels associated with arrest of phagosomal acidification. • Virulence and pathogenesis correlated with the strain's ability to invade ATs (protective intracellular niche).	• Increased expression of efflux pumps. • Reprogramming of host macrophage metabolism (bypassing the IL-1R1 pathway and inducing INF-β). • Remodeling of host transcriptional profiling through epigenetic manipulation. • Altered cytokine/chemokine production in macrophages. • Strains with low PDIM levels associated with low virulence. • Increased adhesion and invasion of ATs: successful dissemination. • Altered levels of CD3, CD4, NKT, CD4/CD8 ratio, TNF levels in serum: immune dysfunction.
Latent *M.tb* infection Granuloma	• Preference toward metabolic pathways using free fatty acids. • Dynamic remodeling of *M.tb* cell proteome and lipid metabolic networks. • Altered exporter function. • DS-*M.tb* becomes phenotypically drug-resistant (cell wall thickening). • Overexpression of efflux pumps.	• Transient drug-tolerance and permanent drug-resistance associated with trehalose-catalytic shift. • Similar phenotype as DS-*M.tb* in the granuloma environment.
Reactivation/active TB disease	• Increased production of mycolic acids, PDIM, SL-1, and PATs. • Biofilm formation in cavities (extracellular matrix of free mycolic acids).	• Currently unknown, although expected to have similarities to the interactions specifically described for DS-*M.tb* strains at this stage of the infection. It is at this stage when DR-*M.tb* is in cavities supposedly exposed to multiple drugs during DR-TB treatment and thus, DR-*M.tb* could also become resistant to these drugs (Sharma et al., 2020).

DS-M.tb, drug-susceptible Mycobacterium tuberculosis; DR-M.tb, drug-resistant Mycobacterium tuberculosis; ALF, alveolar lining fluid; TDM, trehalose dimycolate; ATs, alveolar epithelial cells; H-ALF, high-ALF; L-ALF, low-ALF; TMM, trehalose monomycolate; GLs, glycolipids; PDIM, phthiocerol dimycocerosate; PGL, phenolic glycolipids; PAMPs, pathogen-associated molecular patterns; DCs, dendritic cells.

Thus, *M.tb* exposed to ALF obtained from elderly people and subsequently used to infect human macrophages shows altered trafficking (e.g., decreased phagosome-lysosome fusion) and increased *M.tb* intracellular growth (Moliva et al., 2019). This phenomenon is also observed *in vivo*, when elderly ALF exposed *M.tb* infected mice presented significantly higher bacterial burden and tissue damage in their lungs (Moliva et al., 2019). This increase in virulence when *M.tb* is exposed to elderly ALF is further linked to surfactant protein-D (SP-D) dysfunctionality in the lungs of elderly individuals (Moliva et al., 2019). In this context, polymorphisms in SP-D are linked to TB susceptibility (Azad et al., 2012; Hsieh et al., 2018). SP-D is shown to bind and cluster *M.tb* bacilli, influencing uptake by macrophages and intracellular killing by increased phagosome-lysosome fusion (Ferguson et al., 1999, 2006). These studies demonstrate the importance of the first *M.tb* interactions with soluble host innate immune components and the host ALF composition and functional status in dictating infection progression and disease outcome (Ferguson et al., 1999; Torrelles et al., 2008a). Still, the overall contribution of human ALF as an environmental pressure determinant in the development of DR-TB is a relatively unexplored research area. How *M.tb* changes its metabolism in response to these initial ALF interactions is still largely unknown, as well as which are the specific mycobacterial and host determinants that play a role in establishing a successful infection. Outcomes of *M.tb* infection after interacting with some ALF components [e.g., complement component 3 (C3), mannose binding lectin (MBL), SP-A, SP-D, etc., are described in detail elsewhere (reviewed in Torrelles et al., 2008a; Torrelles and Schlesinger, 2017; Garcia-Vilanova et al., 2019)].

The *M.tb* cell envelope surface composition plays a key role in defining the interactions between *M.tb* and its surroundings, with differences known to exist among *M.tb* clinical isolates (Torrelles et al., 2008b). Indeed, differential glycosylation patterns observed in membrane- and cell envelope-associated *M.tb* proteins, lipids and lipoglycans might contribute to virulence and phenotypic variability across *M.tb* lineages, including drug resistance development (Birhanu et al., 2019). Although little is known about the cell envelope composition of drug-resistant *M.tb*, some studies observed differences in cell envelope thickness between susceptible and resistant strains (**Table 1**). A study using electron transmission microscopy (TEM) showed differences of almost 2 nm and up to 4.6 nm in cell envelope thickness between MDR and XDR strains, respectively, when compared to susceptible isolates, with a denser peptidoglycan layer in XDR-*M.tb* strains (Velayati et al., 2009a) (**Figure 1B**). Further, XXDR-*M.tb* strains in exponential phase studied by atomic force microscopy present alterations in cell morphology (increased roughness and striations, with tubular extensions), probably induced by drug treatment; where a subpopulation of XXDR-*M.tb* bacteria (5–7%) had extraordinarily thick cell envelopes, independent of strain genotype (Velayati et al., 2010). Little information is provided about the drug-resistant strains tested in these studies, and it is still unknown if this thicker cell envelope is a general feature of all DR-*M.tb* strains; however, these microscopy studies suggest that the *M.tb* cell envelope composition is altered in these strains (**Table 1**).

Other studies attempted to analyze differences in lipid profiles between susceptible and DR-*M.tb* strains, as well as to study the role of surface lipids remodeling contributing to the emergence of MDR-*M.tb* phenotypes. Using a high-throughput mass spectrometry-based lipidomic approach, it is shown that MDR-*M.tb* strains have increased levels of free fatty acids, TDM, glycerophospholipids, altered glycerolipids, and unique and distinct lipid signatures when compared to drug-susceptible isolates (Pal et al., 2017) (**Figure 1A** and **Table 1**). Indeed, a potential link between low PDIM and PGL levels in the *M.tb* cell envelope with increased drug susceptibility has been reported using a *tesA* (encoding a type II thioesterase) mutant strain of *Mycobacterium marinum* (Chavadi et al., 2011). On the other hand, a recent study used two genetically distinct *M.tb* clonal pairs (laboratory and clinical drug-sensitive strains and their derived isoniazid-resistant (INHr) mutants) to determine specific *in vitro* changes related to the isoniazid-resistant phenotype through proteomic and lipidomic analyses (Nieto et al., 2018). These INHr isolates presented 26 proteins with altered levels, which were mainly associated with energy metabolism and respiration, but also with lipid metabolism, virulence, adaptation, and cell envelope remodeling. These strains presented activation of an alternative mycolic acid biosynthesis route, which is also observed in DR-*M.tb* strains from other lineages (Singh et al., 2015; Nieto et al., 2016b). Interestingly, of the two INHr mutants studied, only the clinical INHr isolate presented low levels of TDM, TMM (trehalose monomycolate) and PDIM, the latter associated with reduced virulence of this strain (Nieto et al., 2016a). This is in contrast with previous studies, indicating that *M.tb* genetic background may play and important role in determining the remodeling of the *M.tb* cell envelope protein and lipid constituents after acquisition of drug resistance (**Table 1**).

How differently DR-*M.tb* strains interact with the human alveolar environment is still unknown, and it is an active branch of research in our laboratory. In this context, during infection ALF soluble components could modify and shape differently the DR-*M.tb* cell envelope, as well as different DR-*M.tb* strains could have different metabolic responses to the alveolar environment, being an unexplored contributing factor in the development of drug resistance and/or persistence within the host (**Table 1**). Further investigation of a larger number of DR-*M.tb* isolates with resistance to different drugs is needed to support this hypothesis of a generally thickened and altered cell envelope and to determine interactions with ALF components.

Active TB Disease: Interaction With Professional and Non-professional Phagocytes and Other Immune Cells in the Lung

After susceptible or DR-*M.tb* are bathed in immune soluble components of the ALF for an undetermined period of time, AMs are thought to be the first host cellular line of defense against *M.tb*, participating in the pathogen elimination directly or indirectly by activating host innate and adaptive immune responses (**Figure 1C**). AMs are defined as alternative activated, highly phagocytic cells with tightly regulated balance of pro-and

anti-inflammatory responses that avoid the destruction of lung tissue due to excessive inflammation (Evren et al., 2020). *M.tb* has evolved to exploit AM resources as a strategy for host immune evasion and survival (Guirado et al., 2013).

Pathogen-associated molecular patterns (PAMPs) present in the *M.tb* cell envelope are recognized by macrophages and other phagocytic cells like lung submucosal and interstitial dendritic cells (DCs) directly or after being ALF-opsonized through an array of phagocytic receptors [e.g., C-type lectin receptors (CLRs), the mannose receptor (MR), Fc and complement receptors (CRs), DC-specific intercellular adhesion molecule-3-Grabbing non-integrin (DC-SIGN)], and signaling receptors [e.g., Toll-like receptors (TLRs)], among others (Leemans et al., 2003; Sasindran and Torrelles, 2011; Ishikawa et al., 2017; Rajaram et al., 2017; Lugo-Villarino et al., 2018) (**Figure 1C** and **Table 1**). After recognition and binding by a phagocytic receptor, the internalization of *M.tb* will initiate a series of trafficking and signaling cascades and the activation of numerous cellular processes such as *M.tb*-containing phagosome maturation, oxidative and inflammatory responses, antigen processing and presentation, autophagy, and cellular apoptosis, all essential phagocyte killing mechanisms and innate immune responses leading to clearance of *M.tb* infection (Mortaz et al., 2015) (**Figure 1D**). Although there are multiple scenarios of interaction with phagocytic cells via different receptors, it is unclear how *M.tb* uptake through these different phagocytic routes will affect disease progression, or if alternative routes lead to different outcomes.

Nevertheless, *M.tb* has learned to circumvent host defenses (Kang et al., 2005; Singh et al., 2009; Barczak et al., 2017; Cambier et al., 2017; Liu et al., 2018; Peterson et al., 2019), even to limit its exposure to sub-lethal concentrations of antimicrobial drugs potentially promoting the emergence of drug resistance (Wu et al., 2019). Indeed, it is shown that *M.tb* cell envelope surface mimics mammalian mannoproteins by having terminal mannose-containing molecules (e.g., ManLAM, higher-order PIMs, 19 kDa and 24 kDa mannoproteins, PstS-1, Apa, etc.). It is suggested that some strains of *M.tb* use these heavily mannosylated cell surface molecules to interact with lung surfactant proteins such as SP-A (Ragas et al., 2007), and phagocytic receptors including MR (Kang et al., 2005; Torrelles et al., 2006; Torrelles and Schlesinger, 2010; Esparza et al., 2015), Dectin-2 (Decout et al., 2018), and DC-SIGN (Pitarque et al., 2005; Tailleux et al., 2005; Driessen et al., 2009), providing a safe portal of entry for *M.tb* inside alveolar host cells without driving inflammation and limiting tissue pathology (Ehlers, 2010) (**Table 1**). In addition, collectins such as mannose-binding lectin (MBL) promote *M.tb* phagocytosis through bacterial opsonization by binding to ManLAM and PIMs, resulting in establishment of the infection (Denholm et al., 2010; Kalscheuer et al., 2019).

To date, multiple mycobacterial mannoproteins have been identified, with potential roles in virulence, cell invasion, evasion of host defense, and host immunomodulation, reviewed in Mehaffy et al. (2019), Deng et al. (2020), and Tonini et al. (2020). Many of these antigenic glycoproteins contain acyl groups and are defined as lipoglycoproteins (Becker and Sander, 2016). This is the case of the 19 kDa lipoglycoprotein LpqH, which has been shown to have multiple immunomodulatory roles such as: promoting the activation of neutrophils and $CD4^+$ T cells, acting as an adhesin and promoting phagocytosis through lectin receptors, altering macrophage antigen presentation functions, or favoring *M.tb* immune evasion and dissemination by promoting TLR2-dependent macrophage apoptosis, among others (Parra et al., 2017). Another example is the 24 kDa lipoglycoprotein LprG, which binds and transports PIM and ManLAM to the cell surface of *M.tb*, acts as an inhibitor of MHC-II antigen processing in human macrophages *in vitro* (Gehring et al., 2004), although evidence suggests that LprG-reactive T cells are activated through TLR2 and glycosylation of specific MHC-II restriction molecules *in vivo* (Sieling et al., 2008). LprG is also associated with regulating TAG levels, *M.tb* growth rate and virulence in combination with the efflux pump Rv1410 (Martinot et al., 2016). Also, the 38 kDa mannosylated glycoprotein PstS-1 acts as adhesin by binding to the macrophage MR, promoting phagocytosis and intracellular survival (Esparza et al., 2015), and glycosylated lipoprotein SodC (Sartain and Belisle, 2009), a superoxide dismutase produced by *M.tb*, is known to act as a B-cell antigen and contribute to *M.tb* virulence, with a potential role in the defense against the oxidative burst produced *in vivo* inside macrophages (Piddington et al., 2001).

Other groups of glycoproteins involved in *M.tb*–host interactions are the ones belonging to the MCE family, MPT64, or Apa, among others, involved in colonization and invasion of host cells (Sonawane et al., 2012), as well as other proteins such as Rv0227c, HtrA-like serine protease Rv1223, TatA, GlnA1, and the disulfide oxidase DsbA-like enzyme Rv2969c, recently described as mannosylated (Tonini et al., 2020). In addition, GroEL2, a chaperone-like *M.tb* capsule-associated glycoprotein, is shown to contribute to the suboptimal antigen presentation during mycobacterial infection by modulating macrophage and DC proinflammatory responses (Georgieva et al., 2018), supporting the immunogenic role of *M.tb* cell envelope-associated glycoproteins during *M.tb* infection. Other strategies used by *M.tb* to counteract host defense mechanisms are the blockade of phagolysosome biogenesis and phagosomal acidification, where cell envelope components such as ManLAM and TDM among others, play important roles (Fratti et al., 2003; Vergne et al., 2004; Kang et al., 2005) (**Table 1**). A recent study also showed that *M.tb* surface protein Rv1468c binds to ubiquitin, triggering host xenophagy and promoting a controlled persistent intracellular infection while restricting host immune responses (Chai et al., 2019). Increased PIM expression is also reported during *in vitro* macrophage infection, although its involvement in *M.tb* virulence and intracellular survival is still unclear (Glass et al., 2017). The presence of DIM/PDIM lipids in the cell envelope is also associated with arrest of phagosomal acidification and macrophage death (Quigley et al., 2017).

Another important group of cell envelope glycoproteins are the ones associated with drug efflux pumps and mycobacterial cell wall permeability, with key implications in the development of drug resistance and *M.tb* persistence, especially when no genomic mutations are involved [reviewed in Singh et al. (2020)]. A clear example is the MmpL family, a group of inner membrane

proteins associated with the transport of cell envelope lipids such as TMM, PDIM, sulfolipids, and TAG, among others (Melly and Purdy, 2019). Several MmpL proteins also act as drug efflux pumps, e.g., overexpression of MmpL5/S5 and MmpL7 increases mycobacterial resistance to clozafimine, bedaquiline, azole drugs, and isoniazid (Rodrigues et al., 2012; Hartkoorn et al., 2014). Indeed, antibiotic stress induces expression of efflux pumps in clinical *M.tb* strains (Gupta et al., 2010), which can contribute to the emergence of MDR-*M.tb* phenotypes. This is evidenced by the significantly increased expression of multiple efflux pumps, including Tap, among others, in MDR- and XDR-*M.tb* isolates when compared to drug-susceptible strains such as $H_{37}R_v$ (Li et al., 2015; Kardan-Yamchi et al., 2019; Liu et al., 2019) (**Table 1**). Conversely, the outer membrane channel protein CpnT that mediates efficient nutrient uptake during infection, can act as a major drug susceptibility determinant for *M.tb*, where mutations in *cpnT* appear to be associated with increased drug resistance in *M.tb* clinical isolates (Danilchanka et al., 2015). Other transporters involved in drug susceptibility/resistance have been identified in *M.tb* (da Silva et al., 2011; Singh et al., 2020). Overall, a systematic screening of cell surface-exposed lipids and glycoproteins and their functions is key for a better understanding of the pathogenesis and survival strategies adopted by DR-*M.tb* strains, including the development of drug-resistance *in vivo*.

At the end, the balance between macrophage capacity to control intracellular replication and *M.tb* ability to escape macrophage killing may dictate the outcome of the infection. In this delicate interplay, both macrophage phenotypic and functional heterogeneity, as well as *M.tb* genotypic and phenotypic differences displayed by the infecting strain will play important roles in disease progression (Guirado et al., 2013). In this regard, a distinct $CD11c^+$ $CD11b^+$ AM subpopulation with a unique inflammatory signature was shown to preferentially harbor *M.tb* during infection (Lafuse et al., 2019). Conversely, infection with different *M.tb* genotypes and strains induce different cytokine production and immune responses in *in vitro* and *in vivo* models (Lopez et al., 2003; Chacon-Salinas et al., 2005; Sousa et al., 2020). Indeed, there are more than 2,000 genes differentially expressed by macrophages during infection with a lineage 2 MDR-*M.tb* Beijing strain (Leisching et al., 2016). Thus, the biological properties and the metabolic status of infecting *M.tb* bacilli, as well as the surface composition of their cell envelope, will determine how *M.tb* interacts and activates AMs to respond to the infection (Torrelles et al., 2008b; Kalscheuer et al., 2019).

Some evidence indicate that DR-*M.tb* strains may have a different cell envelope thickness and composition than their counterparts, drug-susceptible strains (Velayati et al., 2009a, 2010; Singh et al., 2015; Nieto et al., 2016b, 2018). These potential differences in their cell envelope composition and thickness might be further accentuated by host ALF hydrolases and other host innate immune soluble components, differently impacting the interaction of these strains with AMs and subsequent infection progression. Indeed, MDR-*M.tb* isolates are described to present morphological differences in their cell envelope, that correlate with their selective advantage over other circulating *M.tb* isolates (Silva et al., 2013) (**Table 1**). Several high-throughput transcriptomic, proteomic and lipidomic approaches focused on the cell wall of *rpoB* mutant strains, showing upregulation of enzymes involved in the biosynthesis of PDIM (Bisson et al., 2012), decreased synthesis of fatty acids (du Preez and Loots, 2012), and altered levels of sulfoglycolipids and mycobactin (Lahiri et al., 2016). Other studies indicate that the *rpoB* mutation H526D is associated with altered cell wall physiology and resistance to cell-wall related stresses (Campodonico et al., 2018), and defined *rpoB* mutations as drivers of altered cytokine and chemokine production in macrophages (Ravan et al., 2019) (**Table 1**). Recently, MDR W-Beijing *M.tb* lineage 2 strains with rifampicin-conferring mutations in the *rpoB* gene were shown to overexpress cell envelope lipids such as PDIMs, bypassing the IL-1 receptor type I (IL-1R1) pathway associated with *M.tb* control and inducing INF-β that drives less effective aerobic glycolysis, ultimately reprogramming the metabolism of host macrophages (Howard et al., 2018; Howard and Khader, 2020) (**Figure 1C** and **Table 1**). Still, global effects of *rpoB* mutations on *M.tb* metabolism and cell envelope constitution, and subsequent interactions with host immune cells are still not well characterized, and further studies are needed to determine if the known effects are common to all DR-*M.tb* strains. Finally, there are some clinical evidences indicating that MDR-*M.tb* strains could be remodeling host transcriptional programming through epigenetic manipulation (Marimani et al., 2018; Crimi et al., 2020) (**Table 1**). Indeed, some XDR-*M.tb* strains seem to promote aberrant epigenetic modifications in macrophages, showing increased methylation levels of inflammatory genes in the TLR2 signaling pathway (Behrouzi et al., 2019).

Neutrophils are also key players in the host immune response against *M.tb*, being one of the first recruited innate effector cells to arrive at the infection site, found in large numbers in the lung during active TB disease (Eum et al., 2010) (**Figure 1C**). Upon *M.tb* contact, neutrophils can activate different intracellular and extracellular killing mechanisms to clear the infection, such as phagocytosis, production and release of reactive oxygen intermediates, and secretion of neutrophil extracellular traps (NETs) and granules containing hydrolytic enzymes and antimicrobial peptides (Kroon et al., 2018). Despite neutrophil's protective role against *M.tb*, a regulation of how many neutrophils migrate to the infection site is important. Indeed, increased number of neutrophils in the alveolar space is associated with lung pathology during active TB due to excessive inflammation and tissue destruction, reviewed in detail elsewhere (Dallenga and Schaible, 2016; Muefong and Sutherland, 2020). Furthermore, neutrophil's activity during *M.tb* infection is influenced by the human lung mucosa. In this regard, *M.tb* exposed to ALF enhanced the neutrophil's innate capacity to recognize and kill *M.tb* intracellularly by enhancing phagosome-lysosome fusion events and producing higher levels of TNF and IL-8, while limiting excessive extracellular inflammatory responses and tissue damage by reducing oxidative burst, apoptosis, and degranulation (Arcos et al., 2015). Conversely, released *M.tb* cell envelope fragments after ALF exposure activate neutrophils, leading to an increased local oxidative response

and the production of inflammatory cytokines, which regulate the activity of resting macrophages and thus, contribute to the control of *M.tb* infection (Arcos et al., 2017; Scordo et al., 2017) (**Figure 1C**). Thus, deciphering the effects of drug resistance in the *M.tb*–neutrophil interactions is a research area that needs further investigation.

ATs also play important roles in the early recognition and internalization of *M.tb* in the alveolar compartment (Scordo et al., 2016; Olmo-Fontanez and Torrelles, 2020). Upon *M.tb* recognition, ATs uptake *M.tb* and secrete cytokines, antimicrobial peptides, nitric oxide, surfactant proteins and other soluble components into the ALF, facilitating cell-to-cell crosstalk with other alveolar compartment cells such as AMs and neutrophils, and thus ATs are active players initiating the innate immune response (Garcia-Perez et al., 2003; Boggaram et al., 2013; Chuquimia et al., 2013; Reuschl et al., 2017; Gupta et al., 2018) (**Figure 1C**). Evidence supports that *M.tb* can replicate within ATs type II cells (ATII), localized inside late endosomal vesicles, and in some instances, ATs can process and present *M.tb* antigens via MHC class I, which are efficiently recognized by IFN-γ $CD8^+$ T cells (Harriff et al., 2014), defining also ATs as contributors of the adaptive immune response initiation during *M.tb* infection and TB disease.

Mycobacterium tuberculosis–ATs interactions can occur during: the very first initial stages of primary infection when *M.tb* reaches the alveolar space or after *M.tb* is exposed to soluble host factors present in the lung mucosa (Scordo et al., 2019) (**Figure 1C**), following *M.tb* release from dying necrotic infected cells in advanced disease stages, and/or following release from granulomas during TB reactivation. The role of ATs might differ based on the stage of infection, although most of the studies are focused on initial *M.tb*–AT interactions and little is known about the other potential scenarios. Still, it is postulated that invasion of ATIIs could be beneficial for *M.tb*, as these cells constitute a protective intracellular niche optimal for undetected *M.tb* growth and replication, in part due to *M.tb*-containing late endosomes in ATs that fail to acidify, while evading recognition and clearance by professional alveolar phagocytes (Scordo et al., 2016). Indeed, virulence and pathogenesis of *M.tb* is correlated with their ability to infect ATs (McDonough and Kress, 1995), and differences in the *M.tb* cell envelope composition, predetermined by the nature of host ALF, might be modulating *M.tb*–ATs interactions (**Table 1**). Supporting this, a *M.tb mce1* mutant strain accumulating mycolic acids in its cell envelope can bypass the TLR-2-mediated pro-inflammatory response in ATs (Sequeira et al., 2014). Studies also show that during infection, different *M.tb* genotypic strains induce different immune responses and gene expression patterns and activated pathways in ATs (Mvubu et al., 2016, 2018). Specifically, XDR-*M.tb* strains show increased adhesion and invasion of ATs compared to other genotypes, including drug-susceptible virulent *M.tb* $H_{37}R_v$ and attenuated $H_{37}R_a$ strains, suggesting that successful dissemination of DR-*M.tb* strains might be related to their interaction with the alveolar epithelium (Ashiru et al., 2010) (**Table 1**). We are just beginning to understand the role of ATs during *M.tb* infection and thus, it still remains unclear the differential adaptation of DR-*M.tb* strains to this alveolar cellular environment and its impact in *M.tb* infection and TB disease progression.

Differences in MDR-*M.tb* cell envelope and physiology may lead to different interactions with phagocytic and other immune cells, promoting characteristic host immune responses that might be different than the ones induced by susceptible *M.tb* strains. Indeed, several studies defined an array of different immune responses correlated to DR-*M.tb* strains. For example, a case-control study showed that MDR-TB patients had higher levels of CD3 and CD4 cells, as well as IgM, when compared to drug-susceptible TB patients (Sun et al., 2017). Another study showed that MDR-TB patient's dendritic cells stimulated with MDR-*M.tb* antigens can mediate a significant higher production of IFN-γ by T cells when compared to DCs from uninfected individuals (Hadizadeh Tasbiti et al., 2018), although this study did not address how this MDR-*M.tb* induced immune responses compare to drug-susceptible *M.tb* stimulation. In this regard, another study showed that individuals infected with MDR-*M.tb* Haarlem strains had higher levels of IL-17^+ IFN-γ^- $CD4^+$ T cells through an IL-23 and TGF-β-dependent mechanisms when compared to latently TB-infected and uninfected individuals (Basile et al., 2017). Other host immune correlates have been associated with MDR- and XDR-TB patients when compared to both drug-susceptible TB and uninfected individuals: decreased levels of CD4, CD3/HLA-DR^+ and Fas^+ T cells, and increased levels of NKT and γδ T cells (Kiran et al., 2010); altered CD4/CD8 ratio and higher TNF levels in serum (Song et al., 2018); and low plasma concentrations of human neutrophil peptides (HNP1-3) (Zhu et al., 2011). T-regs might be also contributing to immune dysfunction and *M.tb* persistence in XDR-TB, although it is not clear whether this response is specific to XDR-TB or if it is actually associated with treatment failure independent of the strain's drug resistance profile (Davids et al., 2018).

Overall, these data suggest that DR-*M.tb* strains induce different host immune responses (**Table 1**), and thus; there is a need of more robust clinical studies in order to establish whether the alteration of the host immune responses is an effect or a cause of DR-TB development.

Latent TB: Granuloma Environment

After *M.tb* infection, host innate and adaptive immune responses are mounted, resulting in bacterial clearance, active TB disease or establishment of a latent *M.tb* infection (LTBI), the latter characterized by persistent *M.tb* bacilli contained inside complex structures called granulomas, a highly heterogeneous and dynamic cellular environment in the lung (Ehlers and Schaible, 2012; Guirado and Schlesinger, 2013) (**Figure 1D**). Depending of the granuloma stage (formation, maintenance, maturation, disruption leading to *M.tb* reactivation), these can be studied *in vitro* (Guirado et al., 2015) and *in vivo* using animal models such as mouse, Guinea pigs, rabbit and non-human primate models (NHPs) (Bucsan et al., 2019). Granulomas are composed of immune host cells including AMs, interstitial macrophages, foamy macrophages, monocytes, multi-nucleated giant cells, epithelial-like cells, DCs, NK cells, and neutrophils, surrounded by B and T cells, which tightly control

M.tb replication and dissemination (Torrelles and Schlesinger, 2017; Muefong and Sutherland, 2020) (**Figure 1D**). Host determinants such as TNF, IL-6 and complement are important for cellular recruitment and granuloma maintenance (Dutta and Karakousis, 2014). Further, laser microdissection of individual granulomas, followed by confocal microscopy and proteomics, indicated that the granuloma's center is a pro-inflammatory environment constituted of antimicrobial molecules, ROS and pro-inflammatory eicosanoids, whereas the tissue environment surrounding the granuloma possesses an anti-inflammatory profile (Marakalala et al., 2016).

Indeed, the granuloma environment has been characterized by low pH, hypoxia, oxidative stress, and nutrient starvation, where persisting bacilli are found in a metabolically adapted and non-replicating state (Keren et al., 2011; Eoh et al., 2017; Peterson et al., 2020). Using the *in vitro* granuloma model, *M.tb* is known to undergo metabolic changes, including having a preference toward using free fatty acids as energy source, in detriment of using carbohydrates (Guirado et al., 2015) (**Table 1**). Induction of the DosR regulon is also observed as a contributing factor in response to starvation and adaptation to other environmental stresses, suggesting a relevant role for the DosR regulon on the adaptation and persistence of *M.tb* within granulomas (Mehra et al., 2015; Zheng et al., 2020). Indeed, a proteome-wide study shows a highly dynamic remodeling of the *M.tb* proteome during exponential growth, dormancy and resuscitation (**Table 1**), where the stress-induced DosR regulon contributes 20% to cellular protein content during dormancy (Schubert et al., 2015). Remodeling of the *M.tb* cell envelope composition during dormancy seems to depend on phosphorylation by serine/threonine protein kinases (STPKs), which modulate lipid biosynthetic enzymes through altered exporter function (**Table 1**) (described in detail in Stokas et al., 2020).

Several *in vitro* models of hypoxia/nutrient depletion have also been used to study the cell wall of *M.tb* during forced dormancy [e.g., using media with no detergent or exogenous lipids, or *M.tb* cultured in the Hampshire model chemostat under gradual nutrient depletion in an oxygen-controlled environment (Galagan et al., 2013; Bacon et al., 2014)]. When compared to exponentially growing *M.tb,* dormant *M.tb* has a thicker cell envelope with increased content of TAGs, free mycolic acids, and lipoglycans, in detriment of having less TDMs, TMMs, and PDIMs, but keeping its infectivity and pathogenesis intact in a guinea pig model (Galagan et al., 2013; Bacon et al., 2014). Conversely, *in vitro* lipidomic studies using a modified Wayne's model comparing exponentially grown *M.tb* in normoxic conditions vs. hypoxia-induced dormant *M.tb* (Gopinath et al., 2015) pointed a decrease in mycolic acids production and degradation of cell envelope-associated and cytoplasmic lipids during dormancy, which was attributed to a late onset of dormancy achieved in this model compared to previous studies. This lipid depletion was reversed to normal levels after bacterial re-aeration/resuscitation (Raghunandanan et al., 2019). Based on this model, reduction of mycolic acid content during dormancy could explain why *M.tb* is unable to elicit an appropriate host immune response and clearance during its dormancy state.

Granulomas also provide a protective environment for *M.tb* against anti-TB drugs, and delivery of drugs inside the core of the granuloma becomes challenging (Grobler et al., 2016). *M.tb* adaptation to the granuloma environment is also related to the thickening of the *M.tb* cell envelope (Cunningham and Spreadbury, 1998), where originally drug-susceptible *M.tb* bacillus become phenotypically drug-resistant (Deb et al., 2009; Rodriguez et al., 2014) or persisters (**Figure 1D** and **Table 1**). Further studies also showed *M.tb* overexpressing several efflux pumps within granulomas, contributing to *M.tb* drug tolerance during latency (Adams et al., 2011; Smith et al., 2013) (**Table 1**). In addition, stresses associated to the granuloma microenvironment (inflammation, hypoxia, starvation, among others) are correlated with increased *M.tb* mutation rates and the generation of drug resistance-conferring mutations in the *M.tb* genome, which have important implications in the emergence of DR-*M.tb* phenotypes *in vivo*. For example, host immune-mediated *M.tb* DNA damage suggests a potential role of error-prone DNA repair synthesis in the generation of chromosomally encoded drug resistance mutations, and emergence of persister populations might be associated to an altered metabolic state mediated by the (p)ppGpp alarmone system (Warner and Mizrahi, 2006). Actually, a study based on mathematical modeling estimated that the prevalence of latent MDR-TB is increasing worldwide, where one in every 83 individuals with latent TB is harboring an MDR-*M.tb* strain (Knight et al., 2019). Even though the increased cell envelope thickness observed in DR-*M.tb* strains resembles the appearance of dormant *M.tb* bacilli within granulomas, it is still unclear how DR-*M.tb* strains adapt to the granuloma environment, although transient drug tolerance and permanent drug resistance is associated with the bacterial trehalose-catalytic shift using an *M.tb* persister-like bacilli (PLB) model (Lee et al., 2019) (**Table 1**). Further research is needed in this area in order to develop novel anti-DR-TB strategies.

Finally, other type of host cells associated with granulomas are the mesenchymal stem cells. These host cells are reported to dampen the immune response to *M.tb* infection in tissues surrounding granulomas and thus, can provide a protective intracellular niche for *M.tb* replication during chronic stages of the infection, sheltering *M.tb* bacilli from anti-TB drugs and inflammatory cytokines, via increased PGE2 host signaling (Jain et al., 2020) (**Figure 1D**).

LTBI can persist for decades and thus, it is important to understand the balance between the host cells and *M.tb* creating this immune-controlled environment, especially in the case of DR-TB. When the host immune control is lost due to a variety of intrinsic (e.g., TNF blockade treatments) and external factors (e.g., malnutrition), which are still poorly understood, granulomas fail to contain *M.tb*, and latent infection reactivates and progresses to active TB disease (**Figure 1E**), with the formation of necrotic granulomas, extracellular *M.tb* growth and tissue destruction resulting in lung cavities and *M.tb* dissemination and/or escape to the upper airways to propagate the infection elsewhere. Decreased levels of TNF might be associated with this process, as well as T cell exhaustion or impaired functions (Lin et al., 2007; Garcia-Vilanova et al., 2019), although bacterial and other host factors leading to reactivation

TABLE 2 | Next generation sequencing (NGS) strategies for drug-resistant *Mycobacterium tuberculosis.*

NGS approaches and applications	Advantages	Limitations
WGS • Species/strain identification. • Epidemiological studies. • TB diagnostics. • Molecular determinants of DR-*M.tb.*	• Nucleotide-level resolution: identification of single DR-*M.tb* conferring mutations. • Sequence directly from clinical samples (e.g., sputum). • Reveals *M.tb* genomic diversity and drug resistance microevolution events within the host.	• Only describe the genetic basis of DR-TB. • Limited information regarding DR-*M.tb* adaptation during disease progression and/or interactions with the host immune system.
RNA-Seq • Study DR-*M.tb* responses to drug treatment. • Study DR-*M.tb* metabolic/physiologic changes in different environment/conditions.	• Unbiased whole transcriptome approach. • High-throughput and relatively cost-effective. • Higher sensitivity and specificity compared to other gene expression approaches.	• Unable to distinguish DR-*M.tb* expression profiles of different strains in mixed infections. • Transcriptomic profile of either host or DR-*M.tb* in infected cells/tissues.
Dual RNA-seq • Study transcriptomic changes of both host and DR-*M.tb* simultaneously in infected cells/tissues.	• Global approach to study both host and DR-*M.tb*. • Establish causal host and DR-*M.tb* interactions: genome-wide association studies.	• Needs previous pathogen enrichment, and/or; • Increased sequencing depth to capture bacterial transcriptome (higher costs)
scRNA-seq • Study transcriptional responses of individual host cells or DR-*M.tb* bacillus during different physiological states (e.g., infection and disease progression). • Study DR-*M.tb* persister bacilli subpopulations during drug treatment. • Characterize individual responses in mixed infections.	• Unprecedented level of resolution and technological innovation (ability to barcode transcripts from a single cell). • Study host cell or DR-*M.tb* bacillus population heterogeneity. • Identify rare host cell or DR-*M.tb* bacillus populations. • Better understanding of tissue architecture.	• Lower throughput. • Requires effective isolation of viable *M.tb* bacilli or host cells. • Study either the host cell or *M.tb*, but mostly optimized for eukaryotes. • Need to overcome extra challenges in *M.tb* scRNA-seq [e.g., efficient lysis of the thick *M.tb* cell envelope, capture non-poly(A) mRNA]. • Complex bioinformatics analyses. • High sequencing costs.
Dual scRNA-seq • Study simultaneously host and pathogen at a single-cell resolution.	• Combined advantages of dual and scRNA-seq. • Define spatial-time relationships between host and pathogen. populations during different stages of the infection.	• Same as dual RNA-seq and scRNA-seq. • Difficult to extract enough bacterial information due to the low coverage and DR-*M.tb* bacilli:host ratio. • Complex interpretation of results.
Single-cell multi-Omics • Global understanding of complex DR-*M.tb* and host interactions through the integration of different "Omics" strategies, at a single-cell resolution.	• Systematic approach. • Dissect complex biological networks within a single-cell. • Integrate information from numerous "Omics." • Highest level of information.	• Need for optimized and standardized protocols and bioinformatics approaches. • Difficult to integrate and interpret the data. • Single-cell resolution not easily achievable for some "Omics" approaches.

DR-M.tb, drug-resistant Mycobacterium tuberculosis; scRNA-seq, single-cell RNA sequencing; TB, tuberculosis; WGS, whole-genome sequencing.

remain to be elucidated. It is suggested that during reactivation, *M.tb* increases the production of mycolic acids and other cell envelope components such as PDIM, SL-1, and pentaacyl trehaloses (PATs), increasing cell envelope hydrophobicity and subsequent transmission (Garcia-Vilanova et al., 2019) (**Table 1**). It is also shown that *M.tb* can form biofilms *in vitro* by creating a type of highly hydrophobic extracellular matrix composed of free mycolic acids (Ojha et al., 2008), which might shield the bacterium against host immune responses and high dose of TB drugs in necrotic areas such as cavities in the lungs (Orme, 2011, 2014), allowing bacterial persistence and transmission (**Figure 1E** and **Table 1**).

CONCLUDING REMARKS

The emergence of DR-TB may be worsened by the current COVID-19 pandemic, exacerbating the global health crisis and undermining TB prevention and control strategies. The development of NGS technologies in the past 20 years has revolutionized our understanding of pathogen–host interactions in the infectious disease field, providing genome-wide information in an increasingly time- and cost-effective manner. Thus, whole-genome sequencing (WGS) has been used to gain better insight of *M.tb* biology and TB disease, including epidemiologic investigations, strain identification, TB transmission dynamics, pathogen–host interactions, and as a fast point-of-care (POC) diagnostic test (Wlodarska et al., 2015; Dookie et al., 2018; Brown et al., 2019) (**Table 2**). In this regard, numerous studies demonstrate the value of WGS in predicting DR-*M.tb* phenotypes from the *M.tb* genome sequenced directly from clinical specimens (e.g., sputum) (Groschel et al., 2018; Cohen et al., 2019b) (**Table 2**). Importantly, WGS has provided new insights into the complex underlying molecular mechanisms of *M.tb* drug resistance thanks to its nucleotide-level resolution (Gonzalez-Escalante et al., 2015; Yu et al., 2015; Coll et al., 2018) (**Table 2**). Of interest, studies of *M.tb* clinical isolates collected from the same patient at a single time point (parallel isolates) or at a different time points (serial isolates) during active TB disease progression and/or drug treatment, revealed within-host

M.tb genomic diversity and microevolution events (O'Neill et al., 2015; Ley et al., 2019; Vargas et al., 2020). This observation was associated to *M.tb* drug resistance genes, lipid synthesis and regulation, and host innate immunity regulation. Thus, studying within-host *M.tb* diversity may assist in identifying novel adaptation strategies to drug and immune pressures in the different lung microenvironments (Nimmo et al., 2020).

Importantly, during infection, DR-*M.tb* encounters very different human lung microenvironments, modifying its metabolism to adapt and persist in the human host. Evidence suggests that the initial *M.tb*–host interactions in the alveolar space, such as ALF exposure, play a decisive role in the establishment of *M.tb* infection. Indeed, ALF-derived changes in the *M.tb* cell envelope drive interactions with host innate immune cells, modulating the immune response and infection outcome. Although only a few studies have focused on the *in vitro* and *in vivo* pathogenesis and adaptation of DR-*M.tb* strains, a differential host response to DR-*M.tb* when compared to drug-susceptible strains is reported, which might be due, in part, to an altered *M.tb* envelope composition leading to *M.tb* clearance, active TB disease or latent *M.tb* infection.

In order to develop new strategies to combat the emergence of DR-*M.tb* clinical isolates, it is imperative to decipher how DR-*M.tb* navigates through the different lung tissue microenvironments, adapting its metabolic status to persist. Transcriptomic approaches allow us to better understand the mode of action of particular drugs (**Table 2**), as well as mechanisms of drug tolerance and, in particular, those associated to bacterial metabolic/physiologic changes and to the host immune response, reviewed in detail elsewhere (Liu et al., 2016; Singhania et al., 2018; Briffotaux et al., 2019; Li et al., 2020). In the last decade, dual RNA sequencing has been investigated as an unbiased and global strategy to simultaneously characterize the transcriptional profiling of host and pathogen in infected cells/tissues (Avraham et al., 2016; Westermann et al., 2016, 2017) (**Table 2**). In this regard, dual RNA-Seq of *M. bovis* BCG-infected macrophages showed upregulation of mycobacterial cholesterol degradation and iron acquisition pathways, and recycling of mycolic acids, while host cells upregulated *de novo* cholesterol biosynthesis likely to compensate for the loss of this metabolite due to bacterial catabolism (Rienksma et al., 2015). Further, dual RNA-seq in *M.tb* infected macrophages isolated from mice show that the growth advantage of *M.tb* in AMs compared to interstitial macrophages (IMs) is a direct consequence of the metabolic interplay between *M.tb* and AMs (Pisu et al., 2020). A similar strategy could be applied to study how DR-*M.tb* strains adapt to the lung environment in comparison to drug-susceptible *M.tb* strains, and clearly define transcriptional profiles during infection that could be used in DR-TB diagnostics and/or therapy.

Single-cell RNA sequencing (scRNA-seq) has also gained popularity in recent years as a means to study individual transcriptional responses of different host cell populations during altered physiological states (e.g., infection) at an unprecedented level of resolution (Hosokawa et al., 2017; Kolodziejczyk and Lonnberg, 2018; Vegh and Haniffa, 2018; Hashimoto et al., 2019) (**Table 2**). Thanks to its ability to amplify hundreds to thousands of individual cells in a single preparation, it allows the identification of rare cell populations that otherwise might go undetected in bulk RNA-seq (Tirosh et al., 2016). However, only a few studies have been conducted in the TB field (Gierahn et al., 2017), due to numerous challenges associated with the single-cell isolation and cell wall disruption of *M.tb*, capturing *M.tb* non-polyadenylated mRNA, and amplifying the low amount of prokaryotic starting material [reviewed in detail (Chen et al., 2017; Zhang et al., 2018)] (**Table 2**). Several methods are being developed in order to overcome these microbial scRNA-seq issues (Sheng et al., 2017; Blattman et al., 2020; Imdahl et al., 2020). In this context, scRNA-seq would be critical to study *M.tb* persisters within granulomas, where a small fraction of the total bacterial population is able to survive drug exposure and killing, likely due to different metabolic/physiological state leading to DR-*M.tb*.

Paired dual scRNA-seq, which combines the dual transcriptional profiling of host and pathogen in infected tissues with the high resolution of single-cell analysis, represents the next level in the study of complex host–pathogen interactions (reviewed in Penaranda and Hung, 2019) (**Table 2**). To date, only one study has attempted to simultaneously characterize the transcriptomes of both host and pathogen at a single-cell level in macrophages infected with *Salmonella typhimurium* (Avital et al., 2017), although it was difficult to extract meaningful biological information due to the low bacterial coverage, plus interpretation of results becomes increasingly complex (**Table 2**). Although there is still much to accomplish in this area, this methodology will become invaluable to study the adaptive capabilities of *M.tb* to drug treatments, as well as to study *M.tb*-ALF and *M.tb*-host cell responses during infection. We would be able to respond questions such as why during the same infection some *M.tb* bacterial subpopulations enter drug tolerant states during treatment and others don't, why some *M.tb* bacilli but not others are cleared by the same host cell type during infection, and/or what is the spatial-time relationship between *M.tb* and the host cell defining the infection outcome.

Ultimately, we expect that single-cell multimodal 'Omics' approaches will be the future for dissecting the complex biological networks within each cell, providing a systematic and global picture by analyzing and combining multiple modalities such as genomics, transcriptomics, epigenomics, proteomics, lipidomics, and glycomics, among others (Khan et al., 2019; Zhu et al., 2020) (**Table 2**). Indeed, metabolomics and dual RNA-seq data from *M.tb*-infected macrophages were collected using a multi-omics approach to identify metabolic sub-networks regulated during early *M.tb* infection, showing that *M.tb* consumes up to 33 different host-derived substrates to establish its intracellular niche (Zimmermann et al., 2017).

These novel NGS strategies, including scRNA-seq, dual RNA-seq, and single-cell multi-Omics could provide systematic and global approaches to fully understand the complex DR *M.tb*-host interplay and dynamics during the different stages of the infection, and to uncover regulatory networks critical for DR-*M.tb* survival in the host that could potentially be exploited as novel anti-DR-TB strategies.

AUTHOR CONTRIBUTIONS

AA-G, JIG, and JBT made substantial contributions to the conception, writing, and editing of this review. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Ruth Chapman Cowles and Andrew G. Cowles Memorial Trust to JIG, and the Robert J. Kleberg Jr. and Helen C. Kleberg Foundation, and the National Institutes of Health/National Institute of Allergy and Infectious Diseases (NIH/NIAID) Grant Number RO1 AI-146340 to JBT.

ACKNOWLEDGMENTS

We thank the Texas Biomedical Research Institute family for their efforts to maintain the institute functional during the COVID-19 pandemic.

REFERENCES

Adams, K. N., Takaki, K., Connolly, L. E., Wiedenhoft, H., Winglee, K., Humbert, O., et al. (2011). Drug tolerance in replicating mycobacteria mediated by a macrophage-induced efflux mechanism. *Cell* 145, 39–53. doi: 10.1016/j.cell.2011.02.022

Al-Saeedi, M., and Al-Hajoj, S. (2017). Diversity and evolution of drug resistance mechanisms in *Mycobacterium tuberculosis*. *Infect. Drug Resist.* 10, 333–342. doi: 10.2147/IDR.S144446

Arcos, J., Diangelo, L. E., Scordo, J. M., Sasindran, S. J., Moliva, J. I., Turner, J., et al. (2015). Lung mucosa lining fluid modification of *Mycobacterium tuberculosis* to reprogram human neutrophil killing mechanisms. *J. Infect. Dis.* 212, 948–958. doi: 10.1093/infdis/jiv146

Arcos, J., Sasindran, S. J., Fujiwara, N., Turner, J., Schlesinger, L. S., and Torrelles, J. B. (2011). Human lung hydrolases delineate *Mycobacterium tuberculosis*-macrophage interactions and the capacity to control infection. *J. Immunol.* 187, 372–381. doi: 10.4049/jimmunol.1100823

Arcos, J., Sasindran, S. J., Moliva, J. I., Scordo, J. M., Sidiki, S., Guo, H., et al. (2017). *Mycobacterium tuberculosis* cell wall released fragments by the action of the human lung mucosa modulate macrophages to control infection in an IL-10-dependent manner. *Mucosal Immunol.* 10, 1248–1258. doi: 10.1038/mi.2016.115

Arun, K. B., Madhavan, A., Abraham, B., Balaji, M., Sivakumar, K. C., Nisha, P., et al. (2020). Acetylation of isoniazid is a novel mechanism of isoniazid resistance in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 65:e00456-20. doi: 10.1128/AAC.00456-20

Ashiru, O. T., Pillay, M., and Sturm, A. W. (2010). Adhesion to and invasion of pulmonary epithelial cells by the F15/LAM4/KZN and Beijing strains of *Mycobacterium tuberculosis*. *J. Med. Microbiol.* 59(Pt 5), 528–533. doi: 10.1099/jmm.0.016006-0

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D. T., and Yanai, I. (2017). scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA-sequencing. *Genome Biol.* 18:200. doi: 10.1186/s13059-017-1340-x

Avraham, R., Haseley, N., Fan, A., Bloom-Ackermann, Z., Livny, J., and Hung, D. T. (2016). A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes. *Nat. Protoc.* 11, 1477–1491. doi: 10.1038/nprot.2016.090

Azad, A. K., Sadee, W., and Schlesinger, L. S. (2012). Innate immune gene polymorphisms in tuberculosis. *Infect. Immun.* 80, 3343–3359. doi: 10.1128/IAI.00443-12

Bacon, J., Alderwick, L. J., Allnutt, J. A., Gabasova, E., Watson, R., Hatch, K. A., et al. (2014). Non-replicating *Mycobacterium tuberculosis* elicits a reduced infectivity profile with corresponding modifications to the cell wall and extracellular matrix. *PLoS One* 9:e87329. doi: 10.1371/journal.pone.0087329

Barczak, A. K., Avraham, R., Singh, S., Luo, S. S., Zhang, W. R., Bray, M. A., et al. (2017). Systematic, multiparametric analysis of *Mycobacterium tuberculosis* intracellular infection offers insight into coordinated virulence. *PLoS Pathog.* 13:e1006363. doi: 10.1371/journal.ppat.1006363

Basile, J. I., Kviatcovsky, D., Romero, M. M., Balboa, L., Monteserin, J., Ritacco, V., et al. (2017). *Mycobacterium tuberculosis* multi-drug-resistant strain M induces IL-17(+) IFNgamma(-) CD4(+) T cell expansion through an IL-23 and TGF-beta-dependent mechanism in patients with MDR-TB tuberculosis. *Clin. Exp. Immunol.* 187, 160–173. doi: 10.1111/cei.12873

Becker, K., and Sander, P. (2016). *Mycobacterium tuberculosis* lipoproteins in virulence and immunity - fighting with a double-edged sword. *FEBS Lett.* 590, 3800–3819. doi: 10.1002/1873-3468.12273

Behrouzi, A., Hadifar, S., Amanzadeh, A., Riazi Rad, F., Vaziri, F., and Siadat, S. D. (2019). Aberrant methylation of host macrophages induced by tuberculosis infection. *World J. Microbiol. Biotechnol.* 35:168. doi: 10.1007/s11274-019-2733-7

Bell, L. C. K., and Noursadeghi, M. (2018). Pathogenesis of HIV-1 and *Mycobacterium tuberculosis* co-infection. *Nat. Rev. Microbiol.* 16, 80–90. doi: 10.1038/nrmicro.2017.128

Bellerose, M. M., Baek, S. H., Huang, C. C., Moss, C. E., Koh, E. I., Proulx, M. K., et al. (2019). Common variants in the glycerol kinase gene reduce tuberculosis drug efficacy. *mBio* 10:e663-19. doi: 10.1128/mBio.00663-19

Ben-Kahla, I., and Al-Hajoj, S. (2016). Drug-resistant tuberculosis viewed from bacterial and host genomes. *Int. J. Antimicrob. Agents* 48, 353–360. doi: 10.1016/j.ijantimicag.2016.07.010

Birhanu, A. G., Yimer, S. A., Kalayou, S., Riaz, T., Zegeye, E. D., Holm-Hansen, C., et al. (2019). Ample glycosylation in membrane and cell envelope proteins may explain the phenotypic diversity and virulence in the *Mycobacterium tuberculosis* complex. *Sci. Rep.* 9:2927. doi: 10.1038/s41598-019-39654-9

Bisson, G. P., Mehaffy, C., Broeckling, C., Prenni, J., Rifat, D., Lun, D. S., et al. (2012). Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant *Mycobacterium tuberculosis*. *J. Bacteriol.* 194, 6441–6452. doi: 10.1128/JB.01013-12

Blattman, S. B., Jiang, W., Oikonomou, P., and Tavazoie, S. (2020). Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat. Microbiol.* 5, 1192–1201. doi: 10.1038/s41564-020-0729-6

Boggaram, V., Gottipati, K. R., Wang, X., and Samten, B. (2013). Early secreted antigenic target of 6 kDa (ESAT-6) protein of *Mycobacterium tuberculosis* induces interleukin-8 (IL-8) expression in lung epithelial cells via protein kinase signaling and reactive oxygen species. *J. Biol. Chem.* 288, 25500–25511. doi: 10.1074/jbc.M112.448217

Boldrin, F., Provvedi, R., Cioetto Mazzabo, L., Segafreddo, G., and Manganelli, R. (2020). Tolerance and persistence to drugs: a main challenge in the fight against *Mycobacterium tuberculosis*. *Front. Microbiol.* 11:1924. doi: 10.3389/fmicb.2020.01924

Borrell, S., and Gagneux, S. (2011). Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. *Clin. Microbiol. Infect* 17, 815–820. doi: 10.1111/j.1469-0691.2011.03556.x

Borrell, S., Teo, Y., Giardina, F., Streicher, E. M., Klopper, M., Feldmann, J., et al. (2013). Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evol. Med. Public Health* 2013, 65–74. doi: 10.1093/emph/eot003

Briffotaux, J., Liu, S., and Gicquel, B. (2019). Genome-wide transcriptional responses of *Mycobacterium* to antibiotics. *Front. Microbiol.* 10:249. doi: 10.3389/fmicb.2019.00249

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

Brown, T. S., Challagundla, L., Baugh, E. H., Omar, S. V., Mustaev, A., Auld, S. C., et al. (2019). Pre-detection history of extensively drug-resistant tuberculosis in

KwaZulu-Natal, South Africa. *Proc. Natl. Acad. Sci. U.S.A.* 116, 23284–23291. doi: 10.1073/pnas.1906636116

Bucsan, A. N., Mehra, S., Khader, S. A., and Kaushal, D. (2019). The current state of animal models and genomic approaches towards identifying and validating molecular determinants of *Mycobacterium tuberculosis* infection and tuberculosis disease. *Pathog. Dis.* 77:ftz037. doi: 10.1093/femspd/ftz037

Cambier, C. J., O'Leary, S. M., O'Sullivan, M. P., Keane, J., and Ramakrishnan, L. (2017). Phenolic Glycolipid facilitates mycobacterial escape from microbicidal tissue-resident macrophages. *Immunity* 47, 552.e554–565.e554. doi: 10.1016/j.immuni.2017.08.003

Campodonico, V. L., Rifat, D., Chuang, Y. M., Ioerger, T. R., and Karakousis, P. C. (2018). Altered *Mycobacterium tuberculosis* cell wall metabolism and physiology associated with RpoB mutation H526D. *Front. Microbiol.* 9:494. doi: 10.3389/fmicb.2018.00494

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

Chacon-Salinas, R., Serafin-Lopez, J., Ramos-Payan, R., Mendez-Aragon, P., Hernandez-Pando, R., Van Soolingen, D., et al. (2005). Differential pattern of cytokine expression by macrophages infected in vitro with different *Mycobacterium tuberculosis* genotypes. *Clin. Exp. Immunol.* 140, 443–449. doi: 10.1111/j.1365-2249.2005.02797.x

Chai, Q., Wang, X., Qiang, L., Zhang, Y., Ge, P., Lu, Z., et al. (2019). A *Mycobacterium tuberculosis* surface protein recruits ubiquitin to trigger host xenophagy. *Nat. Commun.* 10:1973. doi: 10.1038/s41467-019-09955-8

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

Chen, Z., Chen, L., and Zhang, W. (2017). Tools for genomic and transcriptomic analysis of microbes at single-cell level. *Front. Microbiol.* 8:1831. doi: 10.3389/fmicb.2017.01831

Chiaradia, L., Lefebvre, C., Parra, J., Marcoux, J., Burlet-Schiltz, O., Etienne, G., et al. (2017). Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane. *Sci. Rep.* 7:12807. doi: 10.1038/s41598-017-12718-4

Chuquimia, O. D., Petursdottir, D. H., Periolo, N., and Fernandez, C. (2013). Alveolar epithelial cells are critical in protection of the respiratory tract by secretion of factors able to modulate the activity of pulmonary macrophages and directly control bacterial growth. *Infect. Immun.* 81, 381–389. doi: 10.1128/IAI.00950-12

Cohen, K. A., Abeel, T., Manson McGuire, A., Desjardins, C. A., Munsamy, V., Shea, T. P., et al. (2015). Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med.* 12:e1001880. doi: 10.1371/journal.pmed.1001880

Cohen, K. A., Manson, A. L., Abeel, T., Desjardins, C. A., Chapman, S. B., Hoffner, S., et al. (2019a). Extensive global movement of multidrug-resistant *M. tuberculosis* strains revealed by whole-genome analysis. *Thorax* 74, 882–889. doi: 10.1136/thoraxjnl-2018-211616

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

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

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., et al. (2011). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* 44, 106–110. doi: 10.1038/ng.1038

Crimi, E., Benincasa, G., Cirri, S., Mutesi, R., Faenza, M., and Napoli, C. (2020). Clinical epigenetics and multidrug-resistant bacterial infections: host remodelling in critical illness. *Epigenetics* 15, 1021–1034. doi: 10.1080/15592294.2020.1748918

Cunningham, A. F., and Spreadbury, C. L. (1998). Mycobacterial stationary phase induced by low oxygen tension: cell wall thickening and localization of the 16-kilodalton alpha-crystallin homolog. *J. Bacteriol.* 180, 801–808. doi: 10.1128/JB.180.4.801-808.1998

da Silva, P. E. A., Von Groll, A., Martin, A., and Palomino, J. C. (2011). Efflux as a mechanism for drug resistance in *Mycobacterium tuberculosis*. *FEMS Immunol. Med. Microbiol.* 63, 1–9. doi: 10.1111/j.1574-695X.2011.00831.x

Dallenga, T., and Schaible, U. E. (2016). Neutrophils in tuberculosis–first line of defence or booster of disease and targets for host-directed therapy? *Pathog. Dis.* 74:ftw012. doi: 10.1093/femspd/ftw012

Danilchanka, O., Pires, D., Anes, E., and Niederweis, M. (2015). The *Mycobacterium tuberculosis* outer membrane channel protein CpnT confers susceptibility to toxic molecules. *Antimicrob. Agents Chemother.* 59, 2328–2336. doi: 10.1128/AAC.04222-14

Davids, M., Pooran, A. S., Pietersen, E., Wainwright, H. C., Binder, A., Warren, R., et al. (2018). Regulatory T cells subvert mycobacterial containment in patients failing extensively drug-resistant tuberculosis treatment. *Am. J. Respir. Crit. Care Med.* 198, 104–116. doi: 10.1164/rccm.201707-1441OC

de Vos, M., Muller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., et al. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. *Antimicrob. Agents Chemother.* 57, 827–832. doi: 10.1128/AAC.01541-12

Deb, C., Lee, C. M., Dubey, V. S., Daniel, J., Abomoelak, B., Sirakova, T. D., et al. (2009). A novel in vitro multiple-stress dormancy model for *Mycobacterium tuberculosis* generates a lipid-loaded, drug-tolerant, dormant pathogen. *PLoS One* 4:e6077. doi: 10.1371/journal.pone.0006077

Decout, A., Silva-Gomes, S., Drocourt, D., Blattes, E., Riviere, M., Prandi, J., et al. (2018). Deciphering the molecular basis of mycobacteria and lipoglycan recognition by the C-type lectin Dectin-2. *Sci. Rep.* 8:16840. doi: 10.1038/s41598-018-35393-5

Deng, G., Zhang, W., Ji, N., Zhai, Y., Shi, X., Liu, X., et al. (2020). Identification of secreted O-mannosylated proteins from BCG and characterization of immunodominant antigens BCG_0470 and BCG_0980. *Front. Microbiol.* 11:407. doi: 10.3389/fmicb.2020.00407

Denholm, J. T., McBryde, E. S., and Eisen, D. P. (2010). Mannose-binding lectin and susceptibility to tuberculosis: a meta-analysis. *Clin. Exp. Immunol.* 162, 84–90. doi: 10.1111/j.1365-2249.2010.04221.x

Dheda, K., Gumbo, T., Maartens, G., Dooley, K. E., McNerney, R., Murray, M., et al. (2017). The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir Med.* doi: 10.1016/S2213-2600(17)30079-6 Online ahead of print

Dong, W., Wang, R., Li, P., Wang, G., Ren, X., Feng, J., et al. (2020). Orphan response regulator Rv3143 increases antibiotic sensitivity by regulating cell wall permeability in *Mycobacterium* smegmatis. *Arch. Biochem. Biophys.* 692:108522. doi: 10.1016/j.abb.2020.108522

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., and Naidoo, K. (2018). Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* 73, 1138–1151. doi: 10.1093/jac/dkx506

Driessen, N. N., Ummels, R., Maaskant, J. J., Gurcha, S. S., Besra, G. S., Ainge, G. D., et al. (2009). Role of phosphatidylinositol mannosides in the interaction between mycobacteria and DC-SIGN. *Infect. Immun.* 77, 4538–4547. doi: 10.1128/Iai.01256-08

du Preez, I., and Loots, D. (2012). Altered fatty acid metabolism due to rifampicin-resistance conferring mutations in the rpoB gene of *Mycobacterium tuberculosis*: mapping the potential of pharmaco-metabolomics for global health and personalized medicine. *OMICS* 16, 596–603. doi: 10.1089/omi.2012.0028

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

Dutta, N. K., and Karakousis, P. C. (2014). Latent tuberculosis infection: myths, models, and molecular mechanisms. *Microbiol. Mol. Biol. Rev.* 78, 343–371. doi: 10.1128/MMBR.00010-14

Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., et al. (2003). Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg. Infect. Dis.* 9, 838–845. doi: 10.3201/eid0907.020803

Ehlers, S. (2010). DC-SIGN and mannosylated surface structures of *Mycobacterium tuberculosis*: a deceptive liaison. *Eur. J. Cell Biol.* 89, 95–101. doi: 10.1016/j.ejcb.2009.10.004

Ehlers, S., and Schaible, U. E. (2012). The granuloma in tuberculosis: dynamics of a host-pathogen collusion. *Front. Immunol.* 3:411. doi: 10.3389/fimmu.2012.00411

Ellner, J. J. (2008). The emergence of extensively drug-resistant tuberculosis: a global health crisis requiring new interventions: part I: the origins and nature of the problem. *Clin. Transl. Sci.* 1, 249–254. doi: 10.1111/j.1752-8062.2008.00060.x

Eoh, H., Wang, Z., Layre, E., Rath, P., Morris, R., Branch Moody, D., et al. (2017). Metabolic anticipation in *Mycobacterium tuberculosis*. *Nat. Microbiol.* 2:17084. doi: 10.1038/nmicrobiol.2017.84

Escribano, I., Rodriguez, J. C., Llorca, B., Garcia-Pachon, E., Ruiz, M., and Royo, G. (2007). Importance of the efflux pump systems in the resistance of *Mycobacterium tuberculosis* to fluoroquinolones and linezolid. *Chemotherapy* 53, 397–401. doi: 10.1159/000109769

Esparza, M., Palomares, B., Garcia, T., Espinosa, P., Zenteno, E., and Mancilla, R. (2015). PstS-1, the 38-kDa *Mycobacterium tuberculosis* glycoprotein, is an adhesin, which binds the macrophage mannose receptor and promotes phagocytosis. *Scand. J. Immunol.* 81, 46–55. doi: 10.1111/sji.12249

Eum, S. Y., Kong, J. H., Hong, M. S., Lee, Y. J., Kim, J. H., Hwang, S. H., et al. (2010). Neutrophils are the predominant infected phagocytic cells in the airways of patients with active pulmonary TB. *Chest* 137, 122–128. doi: 10.1378/chest.09-0903

Evren, E., Ringqvist, E., and Willinger, T. (2020). Origin and ontogeny of lung macrophages: from mice to humans. *Immunology* 160, 126–138. doi: 10.1111/imm.13154

Ferguson, J. S., Martin, J. L., Azad, A. K., McCarthy, T. R., Kang, P. B., Voelker, D. R., et al. (2006). Surfactant protein D increases fusion of *Mycobacterium tuberculosis*-containing phagosomes with lysosomes in human macrophages. *Infect. Immun.* 74, 7005–7009. doi: 10.1128/IAI.01402-06

Ferguson, J. S., Voelker, D. R., McCormack, F. X., and Schlesinger, L. S. (1999). Surfactant protein D binds to *Mycobacterium tuberculosis* bacilli and lipoarabinomannan via carbohydrate-lectin interactions resulting in reduced phagocytosis of the bacteria by macrophages. *J. Immunol.* 163, 312–321.

Fonseca, J. D., Knight, G. M., and McHugh, T. D. (2015). The complex evolution of antibiotic resistance in *Mycobacterium tuberculosis*. *Int. J. Infect. Dis.* 32, 94–100. doi: 10.1016/j.ijid.2015.01.014

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., et al. (2013). *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* 45, 784–790. doi: 10.1038/ng.2656

Fratti, R. A., Chua, J., Vergne, I., and Deretic, V. (2003). *Mycobacterium tuberculosis* glycosylated phosphatidylinositol causes phagosome maturation arrest. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5437–5442. doi: 10.1073/pnas.0737613100

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., et al. (2006a). Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2869–2873. doi: 10.1073/pnas.0511240103

Gagneux, S., Long, C. D., Small, P. M., Van, T., Schoolnik, G. K., and Bohannan, B. J. (2006b). The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. *Science* 312, 1944–1946. doi: 10.1126/science.1124410

Gago, G., Diacovich, L., and Gramajo, H. (2018). Lipid metabolism and its implication in mycobacteria-host interaction. *Curr. Opin. Microbiol.* 41, 36–42. doi: 10.1016/j.mib.2017.11.020

Galagan, J. E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., et al. (2013). The *Mycobacterium tuberculosis* regulatory network and hypoxia. *Nature* 499, 178–183. doi: 10.1038/nature12337

Garcia-Perez, B. E., Mondragon-Flores, R., and Luna-Herrera, J. (2003). Internalization of *Mycobacterium tuberculosis* by macropinocytosis in non-phagocytic cells. *Microb. Pathog.* 35, 49–55. doi: 10.1016/s0882-4010(03)00089-5

Garcia-Vilanova, A., Chan, J., and Torrelles, J. B. (2019). Underestimated manipulative roles of *Mycobacterium tuberculosis* cell envelope glycolipids during infection. *Front. Immunol.* 10:2909. doi: 10.3389/fimmu.2019.02909

Gehring, A. J., Dobos, K. M., Belisle, O. T., Harding, C. V., and Boom, W. H. (2004). *Mycobacterium tuberculosis* LprG (Rv1411c): a novel TLR-2 ligand that inhibits human macrophage class II MHC antigen processing. *J. Immunol.* 173, 2660–2668. doi: 10.4049/jimmunol.173.4.2660

Georgieva, M., Sia, J. K., Bizzell, E., Madan-Lala, R., and Rengarajan, J. (2018). *Mycobacterium tuberculosis* GroEL2 modulates dendritic cell responses. *Infect. Immun.* 86:e387-17. doi: 10.1128/IAI.00387-17

Gierahn, T. M., Wadsworth, M. H. II, Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., et al. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* 14, 395–398. doi: 10.1038/nmeth.4179

Glass, L. N., Swapna, G., Chavadi, S. S., Tufariello, J. M., Mi, K., Drumm, J. E., et al. (2017). *Mycobacterium tuberculosis* universal stress protein Rv2623 interacts with the putative ATP binding cassette (ABC) transporter Rv1747 to regulate mycobacterial growth. *PLoS Pathog.* 13:e1006515. doi: 10.1371/journal.ppat.1006515

Gonzalez-Escalante, L., Penuelas-Urquides, K., Said-Fernandez, S., Silva-Ramirez, B., and Bermudez de Leon, M. (2015). Differential expression of putative drug resistance genes in *Mycobacterium tuberculosis* clinical isolates. *FEMS Microbiol. Lett.* 362:fnv194. doi: 10.1093/femsle/fnv194

Gopinath, V., Raghunandanan, S., Gomez, R. L., Jose, L., Surendran, A., Ramachandran, R., et al. (2015). Profiling the proteome of *Mycobacterium tuberculosis* during dormancy and reactivation. *Mol. Cell. Proteomics* 14, 2160–2176. doi: 10.1074/mcp.M115.051151

Grobler, A., Perez Sierra, Z., and Viljoen, H. J. (2016). Modeling nanoparticle delivery of TB drugs to granulomas. *J. Theor. Biol.* 388, 85–95. doi: 10.1016/j.jtbi.2015.10.004

Groschel, M. I., Walker, T. M., van der Werf, T. S., Lange, C., Niemann, S., and Merker, M. (2018). Pathogen-based precision medicine for drug-resistant tuberculosis. *PLoS Pathog.* 14:e1007297. doi: 10.1371/journal.ppat.1007297

Gu, S., Chen, J., Dobos, K. M., Bradbury, E. M., Belisle, J. T., and Chen, X. (2003). Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. *Mol. Cell. Proteomics* 2, 1284–1296. doi: 10.1074/mcp.M300060-MCP200

Guirado, E., Mbawuike, U., Keiser, T. L., Arcos, J., Azad, A. K., Wang, S. H., et al. (2015). Characterization of host and microbial determinants in individuals with latent tuberculosis infection using a human granuloma model. *mBio* 6:e2537-14. doi: 10.1128/mBio.02537-14

Guirado, E., and Schlesinger, L. S. (2013). Modeling the *Mycobacterium tuberculosis* granuloma - the critical battlefield in host immunity and disease. *Front. Immunol.* 4:98. doi: 10.3389/fimmu.2013.00098

Guirado, E., Schlesinger, L. S., and Kaplan, G. (2013). Macrophages in tuberculosis: friend or foe. *Semin. Immunopathol.* 35, 563–583. doi: 10.1007/s00281-013-0388-2

Gupta, A. K., Katoch, V. M., Chauhan, D. S., Sharma, R., Singh, M., Venkatesan, K., et al. (2010). Microarray analysis of efflux pump genes in multidrug-resistant *Mycobacterium tuberculosis* during stress induced by common anti-tuberculous drugs. *Microb. Drug Resist.* 16, 21–28. doi: 10.1089/mdr.2009.0054

Gupta, N., Kumar, R., and Agrawal, B. (2018). New players in immunity to tuberculosis: the host microbiome, lung epithelium, and innate immune cells. *Front. Immunol.* 9:709. doi: 10.3389/fimmu.2018.00709

Gygli, S. M., Borrell, S., Trauner, A., and Gagneux, S. (2017). Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiol. Rev.* 41, 354–373. doi: 10.1093/femsre/fux011

Hadizadeh Tasbiti, A., Yari, S., Siadat, S. D., Tabarsi, P., Saeedfar, K., and Yari, F. (2018). Cellular immune response in MDR-TB patients to different protein expression of MDR and susceptible *Mycobacterium tuberculosis*: Rv0147, a novel MDR-TB biomarker. *Immunol. Res.* 66, 59–66. doi: 10.1007/s12026-017-8971-6

Harriff, M. J., Cansler, M. E., Toren, K. G., Canfield, E. T., Kwak, S., Gold, M. C., et al. (2014). Human lung epithelial cells contain *Mycobacterium tuberculosis* in a late endosomal vacuole and are efficiently recognized by CD8(+) T cells. *PLoS One* 9:e97515. doi: 10.1371/journal.pone.0097515

Hartkoorn, R. C., Uplekar, S., and Cole, S. T. (2014). Cross-resistance between clofazimine and bedaquiline through upregulation of MmpL5 in

Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* 58, 2979–2981. doi: 10.1128/Aac.00037-14

Hartman, T. E., Wang, Z., Jansen, R. S., Gardete, S., and Rhee, K. Y. (2017). Metabolic perspectives on persistence. *Microbiol. Spectr.* 5:10.1128/microbiolspec.TBTB2-0026-2016. doi: 10.1128/microbiolspec.TBTB2-0026-2016

Hashimoto, K., Kouno, T., Ikawa, T., Hayatsu, N., Miyajima, Y., Yabukami, H., et al. (2019). Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl. Acad. Sci. U.S.A.* 116, 24242–24251. doi: 10.1073/pnas.1907883116

Hermann, C., Karamchand, L., Blackburn, J. M., and Soares, N. C. (2020). Cell envelope proteomics of mycobacteria. *J Proteome Res.* 20, 94–109. doi: 10.1021/acs.jproteome.0c00650

Hosokawa, M., Nishikawa, Y., Kogawa, M., and Takeyama, H. (2017). Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci. Rep.* 7:5199. doi: 10.1038/s41598-017-05436-4

Howard, N. C., and Khader, S. A. (2020). Immunometabolism during *Mycobacterium tuberculosis* Infection. *Trends Microbiol.* 28, 832–850. doi: 10.1016/j.tim.2020.04.010

Howard, N. C., Marin, N. D., Ahmed, M., Rosa, B. A., Martin, J., Bambouskova, M., et al. (2018). *Mycobacterium tuberculosis* carrying a rifampicin drug resistance mutation reprograms macrophage metabolism through cell wall lipid changes. *Nat. Microbiol.* 3, 1099–1108. doi: 10.1038/s41564-018-0245-0

Hsieh, M. H., Ou, C. Y., Hsieh, W. Y., Kao, H. F., Lee, S. W., Wang, J. Y., et al. (2018). Functional analysis of genetic variations in surfactant protein D in mycobacterial infection and their association with tuberculosis. *Front. Immunol.* 9:1543. doi: 10.3389/fimmu.2018.01543

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A. E., and Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat. Microbiol.* 5, 1202–1206. doi: 10.1038/s41564-020-0774-1

Ishikawa, E., Mori, D., and Yamasaki, S. (2017). Recognition of mycobacterial lipids by immune receptors. *Trends Immunol.* 38, 66–76. doi: 10.1016/j.it.2016.10.009

Jain, N., Kalam, H., Singh, L., Sharma, V., Kedia, S., Das, P., et al. (2020). Mesenchymal stem cells offer a drug-tolerant and immune-privileged niche to *Mycobacterium tuberculosis*. *Nat. Commun.* 11:3062. doi: 10.1038/s41467-020-16877-3

Jankute, M., Nataraj, V., Lee, O. Y., Wu, H. H. T., Ridell, M., Garton, N. J., et al. (2017). The role of hydrophobicity in tuberculosis evolution and pathogenicity. *Sci. Rep.* 7:1315. doi: 10.1038/s41598-017-01501-0

Jeanes, C., and O'Grady, J. (2016). Diagnosing tuberculosis in the 21st century - Dawn of a genomics revolution? *Int. J. Mycobacteriol.* 5, 384–391. doi: 10.1016/j.ijmyco.2016.11.028

Jia, X., Yang, L., Dong, M., Chen, S., Lv, L., Cao, D., et al. (2017). The bioinformatics analysis of comparative genomics of *Mycobacterium tuberculosis* complex (MTBC) provides insight into dissimilarities between intraspecific groups differing in host association, virulence, and epitope diversity. *Front. Cell Infect. Microbiol.* 7:88. doi: 10.3389/fcimb.2017.00088

Kalscheuer, R., Palacios, A., Anso, I., Cifuente, J., Anguita, J., Jacobs, W. R. Jr., et al. (2019). The *Mycobacterium tuberculosis* capsule: a cell structure with key implications in pathogenesis. *Biochem. J.* 476, 1995–2016. doi: 10.1042/BCJ20190324

Kang, P. B., Azad, A. K., Torrelles, J. B., Kaufman, T. M., Beharka, A., Tibesar, E., et al. (2005). The human macrophage mannose receptor directs *Mycobacterium tuberculosis* lipoarabinomannan-mediated phagosome biogenesis. *J. Exp. Med.* 202, 987–999. doi: 10.1084/jem.20051239

Kardan-Yamchi, J., Kazemian, H., Haeili, M., Harati, A. A., Amini, S., and Feizabadi, M. M. (2019). Expression analysis of 10 efflux pump genes in multidrug-resistant and extensively drug-resistant *Mycobacterium tuberculosis* clinical isolates. *J. Glob. Antimicrob. Resist.* 17, 201–208. doi: 10.1016/j.jgar.2019.01.003

Keren, I., Minami, S., Rubin, E., and Lewis, K. (2011). Characterization and transcriptome analysis of *Mycobacterium tuberculosis* persisters. *mBio* 2:e100-11. doi: 10.1128/mBio.00100-11

Kester, J. C., and Fortune, S. M. (2014). Persisters and beyond: mechanisms of phenotypic drug resistance and drug tolerance in bacteria. *Crit. Rev. Biochem. Mol. Biol.* 49, 91–101. doi: 10.3109/10409238.2013.869543

Khan, M. M., Ernst, O., Manes, N. P., Oyler, B. L., Fraser, I. D. C., Goodlett, D. R., et al. (2019). Multi-omics strategies uncover host-pathogen interactions. *ACS Infect. Dis.* 5, 493–505. doi: 10.1021/acsinfecdis.9b00080

Kiran, B., Cagatay, T., Clark, P., Kosar, F., Cagatay, P., Yurt, S., et al. (2010). Can immune parameters be used as predictors to distinguish between pulmonary multidrug-resistant and drug-sensitive tuberculosis? *Arch. Med. Sci.* 6, 77–82. doi: 10.5114/aoms.2010.13511

Knight, G. M., McQuaid, C. F., Dodd, P. J., and Houben, R. (2019). Global burden of latent multidrug-resistant tuberculosis: trends and estimates based on mathematical modelling. *Lancet Infect. Dis.* 19, 903–912. doi: 10.1016/S1473-3099(19)30307-X

Koch, A., Mizrahi, V., and Warner, D. F. (2014). The impact of drug resistance on *Mycobacterium tuberculosis* physiology: what can we learn from rifampicin? *Emerg. Microbes Infect.* 3:e17. doi: 10.1038/emi.2014.17

Kolodziejczyk, A. A., and Lonnberg, T. (2018). Global and targeted approaches to single-cell transcriptome characterization. *Brief. Funct. Genomics* 17, 209–219. doi: 10.1093/bfgp/elx025

Kroon, E. E., Coussens, A. K., Kinnear, C., Orlova, M., Moller, M., Seeger, A., et al. (2018). Neutrophils: innate effectors of TB resistance? *Front. Immunol.* 9:2637. doi: 10.3389/fimmu.2018.02637

Lafuse, W. P., Rajaram, M. V. S., Wu, Q., Moliva, J. I., Torrelles, J. B., Turner, J., et al. (2019). Identification of an increased alveolar macrophage subpopulation in old mice that displays unique inflammatory characteristics and is permissive to *Mycobacterium tuberculosis* infection. *J. Immunol.* 203, 2252–2264. doi: 10.4049/jimmunol.1900495

Lahiri, N., Shah, R. R., Layre, E., Young, D., Ford, C., Murray, M. B., et al. (2016). Rifampin resistance mutations are associated with broad chemical remodeling of *Mycobacterium tuberculosis*. *J. Biol. Chem.* 291, 14248–14256. doi: 10.1074/jbc.M116.716704

Lange, C., Chesov, D., Heyckendorf, J., Leung, C. C., Udwadia, Z., and Dheda, K. (2018). Drug-resistant tuberculosis: an update on disease burden, diagnosis and treatment. *Respirology* 23, 656–673. doi: 10.1111/resp.13304

Lange, C., Dheda, K., Chesov, D., Mandalakas, A. M., Udwadia, Z., and Horsburgh, C. R. Jr. (2019). Management of drug-resistant tuberculosis. *Lancet* 394, 953–966. doi: 10.1016/S0140-6736(19)31882-3

Lee, J. J., Lee, S. K., Song, N., Nathan, T. O., Swarts, B. M., Eum, S. Y., et al. (2019). Transient drug-tolerance and permanent drug-resistance rely on the trehalose-catalytic shift in *Mycobacterium tuberculosis*. *Nat. Commun.* 10:2928. doi: 10.1038/s41467-019-10975-7

Leemans, J. C., Florquin, S., Heikens, M., Pals, S. T., van der Neut, R., and van der Poll, T. (2003). CD44 is a macrophage binding site for *Mycobacterium tuberculosis* that mediates macrophage recruitment and protective immunity against tuberculosis. *J. Clin. Invest.* 111, 681–689. doi: 10.1172/Jci200316936

Leisching, G., Pietersen, R. D., Mpongoshe, V., van Heerden, C., van Helden, P., Wiid, I., et al. (2016). The host response to a clinical MDR mycobacterial strain cultured in a detergent-free environment: a global transcriptomics approach. *PLoS One* 11:e0153079. doi: 10.1371/journal.pone.0153079

Ley, S. D., de Vos, M., Van Rie, A., and Warren, R. M. (2019). Deciphering within-host microevolution of *Mycobacterium tuberculosis* through whole-genome sequencing: the phenotypic impact and way forward. *Microbiol. Mol. Biol. Rev.* 83:e00062-18.

Li, G., Zhang, J., Guo, Q., Jiang, Y., Wei, J., Zhao, L. L., et al. (2015). Efflux pump gene expression in multidrug-resistant *Mycobacterium tuberculosis* clinical isolates. *PLoS One* 10:e0119013. doi: 10.1371/journal.pone.0119013

Li, H. C., Guo, H. X., Chen, T., Wang, W., Wu, Z. H., Chen, L., et al. (2020). potential genes related to levofloxacin resistance in *Mycobacterium tuberculosis* based on transcriptome and methylome overlap analysis. *J. Mol. Evol.* 88, 202–209. doi: 10.1007/s00239-019-09926-z

Li, Q. J., Jiao, W. W., Yin, Q. Q., Li, Y. J., Li, J. Q., Xu, F., et al. (2017). Positive epistasis of major low-cost drug resistance mutations rpoB531-TTG and katG315-ACC depends on the phylogenetic background of *Mycobacterium*

tuberculosis strains. *Int. J. Antimicrob. Agents* 49, 757–762. doi: 10.1016/j.ijantimicag.2017.02.009

Lin, P. L., Plessner, H. L., Voitenok, N. N., and Flynn, J. L. (2007). Tumor necrosis factor and tuberculosis. *J. Investig. Dermatol. Symp. Proc.* 12, 22–25. doi: 10.1038/sj.jidsymp.5650027

Liu, F., Chen, J., Wang, P., Li, H., Zhou, Y., Liu, H., et al. (2018). MicroRNA-27a controls the intracellular survival of *Mycobacterium tuberculosis* by regulating calcium-associated autophagy. *Nat. Commun.* 9:4295. doi: 10.1038/s41467-018-06836-4

Liu, J., Shi, W., Zhang, S., Hao, X., Maslov, D. A., Shur, K. V., et al. (2019). Mutations in efflux Pump Rv1258c (Tap) cause resistance to pyrazinamide, isoniazid, and streptomycin in *M. tuberculosis*. *Front. Microbiol.* 10:216. doi: 10.3389/fmicb.2019.00216

Liu, Q. Y., Via, L. E., Luo, T., Liang, L. L., Liu, X., Wu, S. F., et al. (2015). Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. *Sci. Rep.* 5:17507.

Liu, Y. C., Tan, S. M., Huang, L., Abramovitch, R. B., Rohde, K. H., Zimmerman, M. D., et al. (2016). Immune activation of the host cell induces drug tolerance in *Mycobacterium tuberculosis* both in vitro and in vivo. *J. Exp. Med.* 213, 809–825. doi: 10.1084/jem.20151248

Loewenberg, S. (2012). India reports cases of totally drug-resistant tuberculosis. *Lancet* 379:205. doi: 10.1016/s0140-6736(12)60085-3

Lopez, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., Ritacco, V., et al. (2003). A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin. Exp. Immunol.* 133, 30–37. doi: 10.1046/j.1365-2249.2003.02171.x

Lugo-Villarino, G., Troegeler, A., Balboa, L., Lastrucci, C., Duval, C., Mercier, I., et al. (2018). The C-type lectin receptor DC-SIGN has an anti-inflammatory role in human M(IL-4) macrophages in response to *Mycobacterium tuberculosis*. *Front. Immunol.* 9:1123. doi: 10.3389/fimmu.2018.01123

Machado, D., Couto, I., Perdigao, J., Rodrigues, L., Portugal, I., Baptista, P., et al. (2012). Contribution of efflux to the emergence of isoniazid and multidrug resistance in *Mycobacterium tuberculosis*. *PLoS One* 7:e34538. doi: 10.1371/journal.pone.0034538

Maitra, A., Munshi, T., Healy, J., Martin, L. T., Vollmer, W., Keep, N. H., et al. (2019). Cell wall peptidoglycan in *Mycobacterium tuberculosis*: an Achilles' heel for the TB-causing pathogen. *FEMS Microbiol. Rev.* 43, 548–575. doi: 10.1093/femsre/fuz016

Mandal, S., Njikan, S., Kumar, A., Early, J. V., and Parish, T. (2019). The relevance of persisters in tuberculosis drug discovery. *Microbiology* 165, 492–499. doi: 10.1099/mic.0.000760

Manson, A. L., Cohen, K. A., Abeel, T., Desjardins, C. A., Armstrong, D. T., Barry, C. E., et al. (2017). Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* 49, 395–402. doi: 10.1038/ng.3767

Marakalala, M. J., Raju, R. M., Sharma, K., Zhang, Y. J., Eugenin, E. A., Prideaux, B., et al. (2016). Inflammatory signaling in human tuberculosis granulomas is spatially organized. *Nat. Med.* 22, 531–538. doi: 10.1038/nm.4073

Marimani, M., Ahmad, A., and Duse, A. (2018). The role of epigenetics, bacterial and host factors in progression of *Mycobacterium tuberculosis* infection. *Tuberculosis* 113, 200–214. doi: 10.1016/j.tube.2018.10.009

Martinot, A. J., Farrow, M., Bai, L., Layre, E., Cheng, T. Y., Tsai, J. H., et al. (2016). Mycobacterial metabolic syndrome: LprG and Rv1410 regulate triacylglyceride levels, growth rate and virulence in *Mycobacterium tuberculosis*. *PLoS Pathog.* 12:e1005351. doi: 10.1371/journal.ppat.1005351

Mawuenyega, K. G., Forst, C. V., Dobos, K. M., Belisle, J. T., Chen, J., Bradbury, E. M., et al. (2005). *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling. *Mol. Biol. Cell* 16, 396–404. doi: 10.1091/mbc.e04-04-0329

McDonough, K. A., and Kress, Y. (1995). Cytotoxicity for lung epithelial cells is a virulence-associated phenotype of *Mycobacterium tuberculosis*. *Infect. Immun.* 63, 4802–4811. doi: 10.1128/IAI.63.12.4802-4811.1995

Mehaffy, C., Belisle, J. T., and Dobos, K. M. (2019). Mycobacteria and their sweet proteins: an overview of protein glycosylation and lipoglycosylation in *M-tuberculosis*. *Tuberculosis* 115, 1–13. doi: 10.1016/j.tube.2019.01.001

Mehra, S., Foreman, T. W., Didier, P. J., Ahsan, M. H., Hudock, T. A., Kissee, R., et al. (2015). The DosR regulon modulates adaptive immunity and is essential for *Mycobacterium tuberculosis* persistence. *Am. J. Respir. Crit. Care Med.* 191, 1185–1196. doi: 10.1164/rccm.201408-1502OC

Melly, G., and Purdy, G. E. (2019). MmpL proteins in physiology and pathogenesis of *M. tuberculosis*. *Microorganisms* 7:70. doi: 10.3390/microorganisms7030070

Merker, M., Barbier, M., Cox, H., Rasigade, J. P., Feuerriegel, S., Kohl, T. A., et al. (2018). Compensatory evolution drives multidrug-resistant tuberculosis in central Asia. *Elife* 7:e38200. doi: 10.7554/eLife.38200

Middlebrook, G., and Cohn, M. L. (1953). Some observations on the pathogenicity of isoniazid-resistant variants of tubercle Bacilli. *Science* 118, 297–299. doi: 10.1126/science.118.3063.297

Modlin, S. J., Conkle-Gutierrez, D., Kim, C., Mitchell, S. N., Morrissey, C., Weinrick, B. C., et al. (2020). Drivers and sites of diversity in the DNA adenine methylomes of 93 *Mycobacterium tuberculosis* complex clinical isolates. *Elife* 9:e58542. doi: 10.7554/eLife.58542

Moliva, J. I., Duncan, M. A., Olmo-Fontanez, A., Akhter, A., Arnett, E., Scordo, J. M., et al. (2019). The lung mucosa environment in the elderly increases host susceptibility to *Mycobacterium tuberculosis* infection. *J. Infect. Dis.* 220, 514–523. doi: 10.1093/infdis/jiz138

Moliva, J. I., Rajaram, M. V., Sidiki, S., Sasindran, S. J., Guirado, E., Pan, X. J., et al. (2014). Molecular composition of the alveolar lining fluid in the aging lung. *Age* 36:9633. doi: 10.1007/s11357-014-9633-4

Mortaz, E., Adcock, I. M., Tabarsi, P., Masjedi, M. R., Mansouri, D., Velayati, A. A., et al. (2015). Interaction of pattern recognition receptors with *Mycobacterium Tuberculosis*. *J. Clin. Immunol.* 35, 1–10. doi: 10.1007/s10875-014-0103-7

Muefong, C. N., and Sutherland, J. S. (2020). Neutrophils in tuberculosis-associated inflammation and lung pathology. *Front. Immunol.* 11:962. doi: 10.3389/fimmu.2020.00962

Muller, B., Borrell, S., Rose, G., and Gagneux, S. (2013). The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet.* 29, 160–169. doi: 10.1016/j.tig.2012.11.005

Mvubu, N. E., Pillay, B., Gamieldien, J., Bishai, W., and Pillay, M. (2016). Canonical pathways, networks and transcriptional factor regulation by clinical strains of *Mycobacterium tuberculosis* in pulmonary alveolar epithelial cells. *Tuberculosis* 97, 73–85. doi: 10.1016/j.tube.2015.12.002

Mvubu, N. E., Pillay, B., McKinnon, L. R., and Pillay, M. (2018). *Mycobacterium tuberculosis* strains induce strain-specific cytokine and chemokine response in pulmonary epithelial cells. *Cytokine* 104, 53–64. doi: 10.1016/j.cyto.2017.09.027

Namouchi, A., Didelot, X., Schock, U., Gicquel, B., and Rocha, E. P. (2012). After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.* 22, 721–734. doi: 10.1101/gr.129544.111

Nguyen, Q. H., Contamin, L., Nguyen, T. V. A., and Banuls, A. L. (2018). Insights into the processes that drive the evolution of drug resistance in *Mycobacterium tuberculosis*. *Evol. Appl.* 11, 1498–1511. doi: 10.1111/eva.12654

Niederweis, M., Danilchanka, O., Huff, J., Hoffmann, C., and Engelhardt, H. (2010). Mycobacterial outer membranes: in search of proteins. *Trends Microbiol.* 18, 109–116. doi: 10.1016/j.tim.2009.12.005

Nieto, R. L., Mehaffy, C., Creissen, E., Troudt, J., Troy, A., Bielefeldt-Ohmann, H., et al. (2016a). Virulence of *Mycobacterium tuberculosis* after acquisition of isoniazid resistance: individual nature of katG mutants and the possible role of AhpC. *PLoS One* 11:e0166807. doi: 10.1371/journal.pone.0166807

Nieto, R. L., Mehaffy, C., and Dobos, K. M. (2016b). Comparing isogenic strains of Beijing genotype *Mycobacterium tuberculosis* after acquisition of Isoniazid resistance: a proteomics approach. *Proteomics* 16, 1376–1380. doi: 10.1002/pmic.201500403

Nieto, R. L., Mehaffy, C., Islam, M. N., Fitzgerald, B., Belisle, J., Prenni, J., et al. (2018). Biochemical characterization of isoniazid-resistant *Mycobacterium tuberculosis*: can the analysis of clonal strains reveal novel targetable pathways? *Mol. Cell Proteomics* 17, 1685–1701. doi: 10.1074/mcp.RA118.000821

Nimmo, C., Brien, K., Millard, J., Grant, A. D., Padayatchi, N., Pym, A. S., et al. (2020). Dynamics of within-host *Mycobacterium tuberculosis* diversity and heteroresistance during treatment. *EBioMedicine* 55:102747. doi: 10.1016/j.ebiom.2020.102747

Ojha, A. K., Baughn, A. D., Sambandan, D., Hsu, T., Trivelli, X., Guerardel, Y., et al. (2008). Growth of *Mycobacterium tuberculosis* biofilms containing free mycolic acids and harbouring drug-tolerant bacteria. *Mol. Microbiol.* 69, 164–174. doi: 10.1111/j.1365-2958.2008.06274.x

Olmo-Fontanez, A. M., and Torrelles, J. B. (2020). "Alveolar epithelial cells," in *Advances in Host-Directed Therapies Against Tuberculosis*, 1st Edn, eds P. C. Karakousis, R. Hafner, and M. L. Gennaro (Berlin: Springer).

O'Neill, M. B., Mortimer, T. D., and Pepperell, C. S. (2015). Diversity of *Mycobacterium tuberculosis* across evolutionary scales. *PLoS Pathog.* 11:e1005257. doi: 10.1371/journal.ppat.1005257

Orme, I. M. (2011). Development of new vaccines and drugs for TB: limitations and potential strategic errors. *Future Microbiol.* 6, 161–177. doi: 10.2217/fmb.10.168

Orme, I. M. (2014). A new unifying theory of the pathogenesis of tuberculosis. *Tuberculosis* 94, 8–14. doi: 10.1016/j.tube.2013.07.004

Pal, R., Hameed, S., Kumar, P., Singh, S., and Fatima, Z. (2017). Comparative lipidomics of drug sensitive and resistant *Mycobacterium tuberculosis* reveals altered lipid imprints. *3 Biotech* 7:325. doi: 10.1007/s13205-017-0972-6

Parida, S. K., Axelsson-Robertson, R., Rao, M. V., Singh, N., Master, I., Lutckii, A., et al. (2015). Totally drug-resistant tuberculosis and adjunct therapies. *J. Intern. Med.* 277, 388–405. doi: 10.1111/joim.12264

Parra, J., Marcoux, J., Poncin, I., Canaan, S., Herrmann, J. L., Nigou, J., et al. (2017). Scrutiny of *Mycobacterium tuberculosis* 19 kDa antigen proteoforms provides new insights in the lipoglycoprotein biogenesis paradigm. *Sci. Rep.* 7:43682. doi: 10.1038/srep43682

Penaranda, C., and Hung, D. T. (2019). Single-cell RNA sequencing to understand host-pathogen interactions. *ACS Infect Dis.* 5, 336–344. doi: 10.1021/acsinfecdis.8b00369

Peterson, E. J., Bailo, R., Rothchild, A. C., Arrieta-Ortiz, M. L., Kaur, A., Pan, M., et al. (2019). Path-seq identifies an essential mycolate remodeling program for mycobacterial host adaptation. *Mol. Syst. Biol.* 15:e8584. doi: 10.15252/msb.20188584

Peterson, E. J. R., Abidi, A. A., Arrieta-Ortiz, M. L., Aguilar, B., Yurkovich, J. T., Kaur, A., et al. (2020). Intricate genetic programs controlling dormancy in *Mycobacterium tuberculosis*. *Cell Rep.* 31:107577. doi: 10.1016/j.celrep.2020.107577

Phillips, P. C. (2008). Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9, 855–867. doi: 10.1038/nrg2452

Piddington, D. L., Fang, F. C., Laessig, T., Cooper, A. M., Orme, I. M., and Buchmeier, N. A. (2001). Cu,Zn superoxide dismutase of *Mycobacterium tuberculosis* contributes to survival in activated macrophages that are generating an oxidative burst. *Infect. Immun.* 69, 4980–4987. doi: 10.1128/Iai.69.8.4980-4987.2001

Pisu, D., Huang, L., Grenier, J. K., and Russell, D. G. (2020). Dual RNA-seq of Mtb-infected macrophages in vivo reveals ontologically distinct host-pathogen interactions. *Cell Rep.* 30, 335.e334–350.e334. doi: 10.1016/j.celrep.2019.12.033

Pitarque, S., Herrmann, J. L., Duteyrat, J. L., Jackson, M., Stewart, G. R., Lecointe, F., et al. (2005). Deciphering the molecular bases of *Mycobacterium tuberculosis* binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochem. J.* 392(Pt 3), 615–624. doi: 10.1042/BJ20050709

Quigley, J., Hughitt, V. K., Velikovsky, C. A., Mariuzza, R. A., El-Sayed, N. M., and Briken, V. (2017). The cell wall lipid PDIM contributes to phagosomal escape and host cell exit of *Mycobacterium tuberculosis*. *mBio* 8:e148-17. doi: 10.1128/mBio.00148-17

Ragas, A., Roussel, L., Puzo, G., and Riviere, M. (2007). The *Mycobacterium tuberculosis* cell-surface glycoprotein Apa as a potential adhesin to colonize target cells via the innate immune system pulmonary C-type lectin surfactant protein A. *J. Biol. Chem.* 282, 5133–5142. doi: 10.1074/jbc.M610183200

Raghunandanan, S., Jose, L., Gopinath, V., and Kumar, R. A. (2019). Comparative label-free lipidomic analysis of *Mycobacterium tuberculosis* during dormancy and reactivation. *Sci. Rep.* 9:3660. doi: 10.1038/s41598-019-40051-5

Rajaram, M. V. S., Arnett, E., Azad, A. K., Guirado, E., Ni, B., Gerberick, A. D., et al. (2017). *M. tuberculosis*-initiated human mannose receptor signaling regulates macrophage recognition and Vesicle trafficking by FcRgamma-Chain, Grb2, and SHP-1. *Cell Rep.* 21, 126–140. doi: 10.1016/j.celrep.2017.09.034

Ravan, P., Nejad Sattari, T., Siadat, S. D., and Vaziri, F. (2019). Evaluation of the expression of cytokines and chemokines in macrophages in response to rifampin-monoresistant *Mycobacterium tuberculosis* and H37Rv strain. *Cytokine* 115, 127–134. doi: 10.1016/j.cyto.2018.12.004

Reuschl, A. K., Edwards, M. R., Parker, R., Connell, D. W., Hoang, L., Halliday, A., et al. (2017). Innate activation of human primary epithelial cells broadens the host response to *Mycobacterium tuberculosis* in the airways. *PLoS Pathog.* 13:e1006577. doi: 10.1371/journal.ppat.1006577

Rienksma, R. A., Suarez-Diez, M., Mollenkopf, H. J., Dolganov, G. M., Dorhoi, A., Schoolnik, G. K., et al. (2015). Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics* 16:34. doi: 10.1186/s12864-014-1197-2

Rodrigues, L., Machado, D., Couto, I., Amaral, L., and Viveiros, M. (2012). Contribution of efflux activity to isoniazid resistance in the *Mycobacterium tuberculosis* complex. *Infect. Genet. Evol.* 12, 695–700. doi: 10.1016/j.meegid.2011.08.009

Rodriguez, J. G., Hernandez, A. C., Helguera-Repetto, C., Aguilar Ayala, D., Guadarrama-Medina, R., Anzola, J. M., et al. (2014). Global adaptation to a lipid environment triggers the dormancy-related phenotype of *Mycobacterium tuberculosis*. *mBio* 5:e1125-14. doi: 10.1128/mBio.01125-14

Safi, H., Gopal, P., Lingaraju, S., Ma, S., Levine, C., Dartois, V., et al. (2019). Phase variation in *Mycobacterium tuberculosis* glpK produces transiently heritable drug tolerance. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19665–19674. doi: 10.1073/pnas.1907631116

Salverda, M. L., Dellus, E., Gorter, F. A., Debets, A. J., van der Oost, J., Hoekstra, R. F., et al. (2011). Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* 7:e1001321. doi: 10.1371/journal.pgen.1001321

Sartain, M. J., and Belisle, J. T. (2009). N-Terminal clustering of the O-glycosylation sites in the *Mycobacterium tuberculosis* lipoprotein SodC. *Glycobiology* 19, 38–51. doi: 10.1093/glycob/cwn102

Sasindran, S. J., and Torrelles, J. B. (2011). *Mycobacterium Tuberculosis* infection and inflammation: what is beneficial for the host and for the bacterium? *Front. Microbiol.* 2:2. doi: 10.3389/fmicb.2011.00002

Schubert, O. T., Ludwig, C., Kogadeeva, M., Zimmermann, M., Rosenberger, G., Gengenbacher, M., et al. (2015). Absolute proteome composition and dynamics during dormancy and resuscitation of *Mycobacterium tuberculosis*. *Cell Host Microbe* 18, 96–108. doi: 10.1016/j.chom.2015.06.001

Scordo, J. M., Arcos, J., Kelley, H. V., Diangelo, L., Sasindran, S. J., Youngmin, E., et al. (2017). *Mycobacterium tuberculosis* cell wall fragments released upon bacterial contact with the human lung mucosa alter the neutrophil response to infection. *Front. Immunol.* 8:307. doi: 10.3389/fimmu.2017.00307

Scordo, J. M., Knoell, D. L., and Torrelles, J. B. (2016). Alveolar epithelial cells in *Mycobacterium tuberculosis* infection: active players or innocent bystanders? *J. Innate Immun.* 8, 3–14. doi: 10.1159/000439275

Scordo, J. M., Olmo-Fontanez, A. M., Kelley, H. V., Sidiki, S., Arcos, J., Akhter, A., et al. (2019). The human lung mucosa drives differential *Mycobacterium tuberculosis* infection outcome in the alveolar epithelium. *Mucosal Immunol.* 12, 795–804. doi: 10.1038/s41385-019-0156-2

Segura-Cerda, C. A., Lopez-Romero, W., and Flores-Valdez, M. A. (2019). Changes in host response to *Mycobacterium tuberculosis* infection associated with type 2 diabetes: beyond hyperglycemia. *Front. Cell Infect. Microbiol.* 9:342. doi: 10.3389/fcimb.2019.00342

Sequeira, P. C., Senaratne, R. H., and Riley, L. W. (2014). Inhibition of toll-like receptor 2 (TLR-2)-mediated response in human alveolar epithelial cells by mycolic acids and *Mycobacterium tuberculosis* mce1 operon mutant. *Pathog. Dis.* 70, 132–140. doi: 10.1111/2049-632X.12110

Sharma, D., Sharma, S., and Sharma, J. (2020). Potential strategies for the management of drug-resistant tuberculosis. *J. Glob. Antimicrob. Resist.* 22, 210–214. doi: 10.1016/j.jgar.2020.02.029

Shcherbakov, D., Akbergenov, R., Matt, T., Sander, P., Andersson, D. I., and Bottger, E. C. (2010). Directed mutagenesis of *Mycobacterium* smegmatis 16S rRNA to reconstruct the in vivo evolution of aminoglycoside resistance in

Mycobacterium tuberculosis. *Mol. Microbiol.* 77, 830–840. doi: 10.1111/j.1365-2958.2010.07218.x

Sheng, K. W., Cao, W. J., Niu, Y. C., Deng, Q., and Zong, C. H. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* 14, 267–270. doi: 10.1038/Nmeth.4145

Sherman, D. R., Mdluli, K., Hickey, M. J., Arain, T. M., Morris, S. L., Barry, C. E., et al. (1996). Compensatory ahpC gene expression in isoniazid-resistant *Mycobacterium tuberculosis*. *Science* 272, 1641–1643. doi: 10.1126/science.272.5268.1641

Sieling, P. A., Hill, P. J., Dobos, K. M., Brookman, K., Kuhlman, A. M., Fabri, M., et al. (2008). Conserved mycobacterial lipoglycoproteins activate TLR2 but also require glycosylation for MHC class II-Restricted T cell activation. *J. Immunol.* 180, 5833–5842. doi: 10.4049/jimmunol.180.9.5833

Silva, C., Perdigao, J., Alverca, E., de Matos, A. P. A., Carvalho, P. A., Portugal, I., et al. (2013). Exploring the contribution of mycobacteria characteristics in their interaction with human macrophages. *Microsc. Microanal.* 19, 1159–1169. doi: 10.1017/S1431927613001906

Singh, A., Crossman, D. K., Mai, D., Guidry, L., Voskuil, M. I., Renfrow, M. B., et al. (2009). *Mycobacterium tuberculosis* WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathog.* 5:e1000545. doi: 10.1371/journal.ppat.1000545

Singh, A., Gopinath, K., Sharma, P., Bisht, D., Sharma, P., Singh, N., et al. (2015). Comparative proteomic analysis of sequential isolates of *Mycobacterium tuberculosis* from a patient with pulmonary tuberculosis turning from drug sensitive to multidrug resistant. *Indian J. Med. Res.* 141, 27–45. doi: 10.4103/0971-5916.154492

Singh, R., Dwivedi, S. P., Gaharwar, U. S., Meena, R., Rajamani, P., and Prasad, T. (2020). Recent updates on drug resistance in *Mycobacterium tuberculosis*. *J. Appl. Microbiol.* 128, 1547–1567. doi: 10.1111/jam.14478

Singhania, A., Wilkinson, R. J., Rodrigue, M., Haldar, P., and O'Garra, A. (2018). The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis. *Nat. Immunol.* 19, 1159–1168. doi: 10.1038/s41590-018-0225-9

Smith, T., Wolff, K. A., and Nguyen, L. (2013). Molecular biology of drug resistance in *Mycobacterium tuberculosis*. *Curr. Top. Microbiol. Immunol.* 374, 53–80. doi: 10.1007/82_2012_279

Sonawane, A., Mohanty, S., Jagannathan, L., Bekolay, A., and Banerjee, S. (2012). Role of glycans and glycoproteins in disease development by *Mycobacterium tuberculosis*. *Crit. Rev. Microbiol.* 38, 250–266. doi: 10.3109/1040841X.2011.653550

Song, T., Li, L. J., Liu, J. L., and Geng, S. J. (2018). Study on the relationship between changes of immune cells and TNF-alpha in peripheral blood of patients with multidrug-resistant and extensively drug-resistant tuberculosis. *Eur. Rev. Med. Pharmacol. Sci.* 22, 1061–1065. doi: 10.26355/eurrev_201802_14389

Sousa, J., Ca, B., Maceiras, A. R., Simoes-Costa, L., Fonseca, K. L., Fernandes, A. I., et al. (2020). *Mycobacterium tuberculosis* associated with severe tuberculosis evades cytosolic surveillance systems and modulates IL-1beta production. *Nat. Commun.* 11:1949. doi: 10.1038/s41467-020-15832-6

Stokas, H., Rhodes, H. L., and Purdy, G. E. (2020). Modulation of the *M. tuberculosis* cell envelope between replicating and non-replicating persistent bacteria. *Tuberculosis* 125:102007. doi: 10.1016/j.tube.2020.102007

Sun, E. T., Xia, D., Li, B. H., Ma, J., Dong, Y. Y., Ding, S. S., et al. (2017). Association of immune factors with drug-resistant tuberculosis: a case-control study. *Med. Sci. Monit.* 23, 5330–5336. doi: 10.12659/msm.904309

Swain, S. S., Sharma, D., Hussain, T., and Pati, S. (2020). Molecular mechanisms of underlying genetic factors and associated mutations for drug resistance in *Mycobacterium tuberculosis*. *Emerg. Microbes Infect.* 9, 1651–1663. doi: 10.1080/22221751.2020.1785334

Tailleux, L., Pham-Thi, N., Bergeron-Lafaurie, A., Herrmann, J. L., Charles, P., Schwartz, O., et al. (2005). DC-SIGN induction in alveolar macrophages defines privileged target host cells for mycobacteria in patients with tuberculosis. *PLoS Med.* 2:e381. doi: 10.1371/journal.pmed.0020381

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H. II, Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Tonini, L., Sadet, B., Stella, A., Bouyssie, D., Nigou, J., Burlet-Schiltz, O., et al. (2020). Potential plasticity of the mannoprotein repertoire associated to *Mycobacterium tuberculosis* virulence unveiled by mass spectrometry-based glycoproteomics. *Molecules* 25:2348. doi: 10.3390/molecules25102348

Torrelles, J. B., Azad, A. K., Henning, L. N., Carlson, T. K., and Schlesinger, L. S. (2008a). Role of C-type lectins in mycobacterial infections. *Curr. Drug Targets* 9, 102–112. doi: 10.2174/138945008783502467

Torrelles, J. B., Knaup, R., Kolareth, A., Slepushkina, T., Kaufman, T. M., Kang, P., et al. (2008b). Identification of *Mycobacterium tuberculosis* clinical isolates with altered phagocytosis by human macrophages due to a truncated lipoarabinomannan. *J. Biol. Chem.* 283, 31417–31428. doi: 10.1074/jbc.M806350200

Torrelles, J. B., Azad, A. K., and Schlesinger, L. S. (2006). Fine discrimination in the recognition of individual species of phosphatidyl-myo-inositol mannosides from *Mycobacterium tuberculosis* by C-type lectin pattern recognition receptors. *J. Immunol.* 177, 1805–1816. doi: 10.4049/jimmunol.177.3.1805

Torrelles, J. B., and Schlesinger, L. S. (2010). Diversity in *Mycobacterium tuberculosis* mannosylated cell wall determinants impacts adaptation to the host. *Tuberculosis* 90, 84–93. doi: 10.1016/j.tube.2010.02.003

Torrelles, J. B., and Schlesinger, L. S. (2017). Integrating lung physiology, immunology, and tuberculosis. *Trends Microbiol.* 25, 688–697. doi: 10.1016/j.tim.2017.03.007

Vargas, R. Jr., and Farhat, M. R. (2020). Antibiotic treatment and selection for glpK mutations in patients with active tuberculosis disease. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3910–3912. doi: 10.1073/pnas.1920788117

Vargas, R. Jr., Freschi, L., Marin, M., Epperson, L. E., Smith, M., Oussenko, I., et al. (2020). In-host population dynamics of *M. tuberculosis* during treatment failure. *bioRxiv [Preprint]* doi: 10.1101/726430

Vegh, P., and Haniffa, M. (2018). The impact of single-cell RNA sequencing on understanding the functional organization of the immune system. *Brief. Funct. Genomics* 17, 265–272. doi: 10.1093/bfgp/ely003

Velayati, A. A., Farnia, P., Ibrahim, T. A., Haroun, R. Z., Kuan, H. O., Ghanavi, J., et al. (2009a). Differences in cell wall thickness between resistant and nonresistant strains of *Mycobacterium tuberculosis*: using transmission electron microscopy. *Chemotherapy* 55, 303–307. doi: 10.1159/000226425

Velayati, A. A., Masjedi, M. R., Farnia, P., Tabarsi, P., Ghanavi, J., ZiaZarifi, A. H., et al. (2009b). Emergence of new forms of totally drug-resistant tuberculosis bacilli super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. *Chest* 136, 420–425. doi: 10.1378/chest.08-2427

Velayati, A. A., Farnia, P., and Masjedi, M. R. (2013). The totally drug resistant tuberculosis (TDR-TB). *Int. J. Clin. Exp. Med.* 6, 307–309.

Velayati, A. A., Farnia, P., Merza, M. A., Zhavnerko, G. K., Tabarsi, P., Titov, L. P., et al. (2010). New insight into extremely drug-resistant tuberculosis: using atomic force microscopy. *Eur. Respir. J.* 36, 1490–1493. doi: 10.1183/09031936.00064510

Vergne, I., Fratti, R. A., Hill, P. J., Chua, J., Belisle, J., and Deretic, V. (2004). *Mycobacterium tuberculosis* phagosome maturation arrest: mycobacterial phosphatidylinositol analog phosphatidylinositol mannoside stimulates early endosomal fusion. *Mol. Biol. Cell* 15, 751–760. doi: 10.1091/mbc.e03-05-0307

Vilcheze, C., and Jacobs, W. R. Jr. (2019). The isoniazid paradigm of killing, resistance, and persistence in *Mycobacterium tuberculosis*. *J. Mol. Biol.* 431, 3450–3461. doi: 10.1016/j.jmb.2019.02.016

Warner, D. F., and Mizrahi, V. (2006). Tuberculosis chemotherapy: the influence of bacillary stress and damage response pathways on drug efficacy. *Clin. Microbiol. Rev.* 19, 558–570. doi: 10.1128/CMR.00060-05

Westermann, A. J., Barquist, L., and Vogel, J. (2017). Resolving host-pathogen interactions by dual RNA-seq. *PLoS Pathog.* 13:e1006033. doi: 10.1371/journal.ppat.1006033

Westermann, A. J., Forstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., et al. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature* 529, 496–501. doi: 10.1038/nature16547

WHO (2019). *Global Tuberculosis Report 2019*. Geneva: WHO.

Wlodarska, M., Johnston, J. C., Gardy, J. L., and Tang, P. (2015). A microbiological revolution meets an ancient disease: improving the management of tuberculosis with genomics. *Clin. Microbiol. Rev.* 28, 523–539. doi: 10.1128/CMR.00124-14

Wu, Q., Hossfeld, A., Gerberick, A., Saljoughian, N., Tiwari, C., Mehra, S., et al. (2019). Effect of *Mycobacterium tuberculosis* enhancement of macrophage P-glycoprotein expression and activity on intracellular survival during antituberculosis drug treatment. *J. Infect. Dis.* 220, 1989–1998. doi: 10.1093/infdis/jiz405

Yu, G. H., Cui, Z. L., Sun, X., Peng, J. F., Jiang, J., Wu, W., et al. (2015). Gene expression analysis of two extensively drug-resistant tuberculosis isolates show that two-component response systems enhance drug resistance. *Tuberculosis* 95, 303–314. doi: 10.1016/j.tube.2015.03.008

Zaman, K. (2010). Tuberculosis: a global health problem. *J. Health Popul. Nutr.* 28, 111–113. doi: 10.3329/jhpn.v28i2.4879

Zhang, Y., Gao, J., Huang, Y., and Wang, J. (2018). Recent developments in single-Cell RNA-Seq of microorganisms. *Biophys. J.* 115, 173–180. doi: 10.1016/j.bpj.2018.06.008

Zhang, Y., Yew, W. W., and Barer, M. R. (2012). Targeting persisters for tuberculosis control. *Antimicrob. Agents Chemother.* 56, 2223–2230. doi: 10.1128/AAC.06288-11

Zheng, H., Williams, J. T., Aleiwi, B., Ellsworth, E., and Abramovitch, R. B. (2020). Inhibiting *Mycobacterium tuberculosis* DosRST signaling by targeting response regulator DNA binding and sensor kinase heme. *ACS Chem. Biol.* 15, 52–62. doi: 10.1021/acschembio.8b00849

Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. *Nat. Methods* 17, 11–14. doi: 10.1038/s41592-019-0691-5

Zhu, L. M., Liu, C. H., Chen, P., Dai, A. G., Li, C. X., Xiao, K., et al. (2011). Multidrug-resistant tuberculosis is associated with low plasma concentrations of human neutrophil peptides 1-3. *Int. J. Tuberc. Lung Dis.* 15, 369–374.

Zimmermann, M., Kogadeeva, M., Gengenbacher, M., McEwen, G., Mollenkopf, H. J., Zamboni, N., et al. (2017). Integration of metabolomics and transcriptomics reveals a complex diet of *Mycobacterium tuberculosis* during Early macrophage infection. *mSystems* 2:e57-17. doi: 10.1128/mSystems.00057-17

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.

Copyright © 2021 Allué-Guardia, García and Torrelles. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 February 2021
doi: 10.3389/fmicb.2020.623574

Predominance of Non-carbapenemase Producing Carbapenem-Resistant Enterobacterales in South Texas

*Cody A. Black[1,2,3], Wonhee So[4], Steven S. Dallas[2,3], Gerard Gawrys[1,5], Raymond Benavides[1,2], Samantha Aguilar[1,2,3], Chang-Jui Chen[1,2], James F. Shurko[1,2] and Grace C. Lee[1,2]**

[1] The University of Texas at Austin, College of Pharmacy, Austin, TX, United States, [2] School of Medicine, The University of Texas Health San Antonio, San Antonio, TX, United States, [3] Department of Pathology and Laboratory Medicine, The University of Texas Health San Antonio, University Health System, San Antonio, TX, United States, [4] School of Pharmacy, Long Island University, Brooklyn, US and Pharmacy, New York Presbyterian/Weill Cornell Medical Center, New York, NY, United States, [5] Methodist Healthcare System, San Antonio, TX, United States

OPEN ACCESS

Edited by:
Jianying Gu,
College of Staten Island,
United States

Reviewed by:
Krisztina M. Papp-Wallace,
Louis Stokes Cleveland VA Medical Center, United States
Anusak Kerdsin,
Kasetsart University, Thailand

***Correspondence:**
Grace C. Lee
leeg3@uthscsa.edu

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

Received: 30 October 2020
Accepted: 30 December 2020
Published: 10 February 2021

Citation:
Black CA, So W, Dallas SS, Gawrys G, Benavides R, Aguilar S, Chen C-J, Shurko JF and Lee GC (2021) Predominance of Non-carbapenemase Producing Carbapenem-Resistant Enterobacterales in South Texas. Front. Microbiol. 11:623574. doi: 10.3389/fmicb.2020.623574

Background: Carbapenem-resistant Enterobacterales (CRE) pose a significant global public health threat. Resistance among CRE is particularly complex, owing to numerous possible resistance mechanisms and broad definitions. We aimed to characterize the clinical and molecular profiles of CRE in the South Texas region.

Materials and methods: We compared the clinical, genotypic, and phenotypic profiles of carbapenemase producing Enterobacterales (CPE) with those of non-carbapenemase producers (NCPE) isolated from South Texas, United States between 2011 and 2019. Molecular characteristics and resistance mechanisms were analyzed using whole-genome sequences.

Results: The majority (59%) of the CRE isolates were NCPE while 41% of isolates harbored carbapenemases, predmonantly bla_{KPC}-type. The most common CPE was *Klebsiella pneumoniae* while majority of *Enterobacter cloacae* and *Escherichia coli* were NCPE Among *K. pneumoniae,* the clonal group 307 has emerged as a predmoninant group and was associated with as many CRE infections as the previous common clonal group 258. Patients with NCPE compared to CPE infections were associated with higher antimicrobial exposure prior to culture collection (days of therapy, 795 vs. 242; $p < 0.001$) and emergency department visits within past 90 days (22% vs. 4%; $p = 0.011$). The all cause 30-day mortality was 21%.

Conclusions: This study highlights the diversity of resistance mechanisms underlying CRE in South Texas, with 59% not harboring a carbapenemase. Individuals with NCPE infections were more likely to have had prior antimicrobial therapy and emergency department visits compared to those with CPE. Identification and distinction of these mechanisms by rapid identification of species and carbapenemase would allow for optimal treatment and infection control efforts.

Keywords: carbapenem, resistance, epidemiology, outcomes, Enterobacterales

INTRODUCTION

Carbapenem-resistant Enterobacterales(CRE), formerly known as Enterobacteriaceae, are a group of organisms that are among the most difficult to treat. It was given a level one priority by the World Health Organization in 2017 (WHO, 2020), and also remains one of the top five most urgent public health threat in 2019 by U.S. Centers of Disease Control and Prevention (Centers for Disease Control and Prevention (U.S.), 2019).

The broad definition of "CRE" remains ambiguous. For example, CREs can include those due to porin-deficient Enterobacterales, Enterobacterales with intrinsic imipenem resistance, and Enterobacterales with carbapenemase enzymes that remain "susceptible" to carbapenems. Globally, common carbapenemases in Enterobacterales include the *Klebsiella pneumoniae* carbapenemases (KPC), oxacillinase (OXA)-48-like β-lactamases, and metallo-β-lactamases, such as New-Delhi-metallo-β-lactamases (NDM), the active-in-imipenem family of carbapenemases, and Verona integron-encoded metallo-β-lactamases (VIM). In the U.S., KPC is the most prevalent type, however; sporadic outbreaks of class B (e.g., VIM, NDM, IMP) or class D (e.g., OXA-48) β-lactamase infections have also occurred (Chea et al., 2015). Importantly, emerging reports indicate that a subset of CDC-defined CRE in the United States do not produce carbapenemases (Duin et al., 2020). Specifically, carbapenem resistance *in K. pneumoniae* can occur from the production of extended-spectrum β-lactamases (ESBLs) and/or AmpC enzymes in combination with decreased antibiotic intracellular accumulation by mutations in outer membrane proteins OmpK35 and OmpK36 or over-expression of efflux pumps.

Treatment of CRE infections remains challenging for clinicians since it encompasses a wide range of genera, different resistant mechanisms requiring dissimilar treatment options, and variable local and global epidemiology. Distinguishing between these different mechanisms, as well as identification of the carbapenemase type, warrants further investigation as these have important treatment implications, particularly for newer β-lactam/β-lactamase inhibitor combinations. Herein, we characterized clinical, genotypic, and phenotypic profiles of sequentially collected CRE isolates from South Texas, United States to investigate the epidemiologic characteristics. We compared mechanisms and implications among Enterobacterales that are carbapenemase producers (CPE) with those that are non-carbapenemase producers (NCPE).

RESULTS

Microbiological Characteristics

We identified 99 CRE isolates collected from 85 patients. Most of the infections were caused by *K. pneumoniae* (43%), followed by *Escherichia coli* (27%), *Enterobacter cloacae* (12%), *Klebsiella aerogenes* (7%), *Citrobacter freundii* (3%), *Serratia marcescens* (3%), *Klebsiella michiganensis* (2%), and *Morganella morganii* (2%), and *Hafnia alvei* (1%). Carbapenemase genes were present in 42 (41%) of CRE, referred to as CPE. In 58 (59%), *in vitro* resistance to at least one carbapenem was confirmed in the absence of any carbapenemase gene, referred as NCPE. There were 17 isolates that were unconfirmed-CRE (identified initially as carbapenem-resistant by local laboratories, but were found to be susceptible or intermediate to all tested carbapenems upon repeat testing) (**Table 1**).

The most common source of CRE isolates was urine [35 (36%)], followed by respiratory [15 (15%)], wound tissue/bone [11 (11%)], and blood [8 (8%)]. Individuals with CRE infections with NCPE were more likely to have respiratory infections compared with patients with CPE ($p = 0.044$). There were similar distributions of CPE and NCPE associated with urinary, blood, and wound infections (**Table 2**). Distribution by species are further described in **Supplementary Table 1**.

Molecular and Antimicrobial Susceptibility Characteristics

Resistance mechanisms were evaluated for CRE isolates (**Table 1**). KPC-type enzymes were the most common carbapenemases and were identified in 30 isolates (30.3%). Observed carbapenemase genes included bla_{KPC-2} [21 (21%) of 99], bla_{KPC-3} [8 (8%)], bla_{NDM} [5 (5%)], and $bla_{OXA-232}$-like [2 (2%); **Table 1**]. One *S. marcescens* isolate was found to simultaneously carry bla_{KPC-2} and bla_{SME-3}. Additionally, it harbored bla_{SRT-2}. Extended spectrum β-lactamase genes were commonly found in CPE including bla_{CTX-M} [17 (42%)] and bla_{SHV} [9 (22%) of 41]. NCPE isolates predominantly carried bla_{CTX-M} [28 (48%) of 59]. Among NCPE, there were no differences in the rate of bla_{AmpC} carriage compared to CPEs [12 (20%) vs. 5 (12%); $p = 0.15$] and unconfirmed CRE [4 (20%); $p = 0.21$]. Among *K. pneumoniae* isolates, mutations in genes encoding outer membrane porins OmpK35 and OmpK36 were present in 62% of NCPE and 53% of unconfirmed CRE. The main mutations included premature stop codons and Gly115-Asp116 insertion in the gene encoding OmpK36. The majority of *Klebsiella* contained several mutations in OmpK37. Specifically, I70M, I128M, N230G, M233Q, T234H, Q235Y, N237H, R239K, E244D, N274S, D275T, and V277I were observed, presumably slowing the permeation of β-lactams through the channel. However, these implications are less defined.

Antimicrobial susceptibility results are shown in **Table 3**. Among the non-β-lactam-inhibitor antimicrobial agents that serve as candidates for the treatment of CRE, tigecycline, amikacin, and colistin showed the highest rates of susceptibility of 97, 91, and 89% respectively. Comparatively, CPE maintained higher susceptibility to older generation tetracyclines (doxycycline and minocycline) compared to NCPE (69 and 89% vs 41 and 52%, respectively); whereas, NCPE isolates maintained higher susceptibility to cefepime compared to CPE (41% vs. 19%). Polymyxin resistance was detected in two *K. pneumoniae*, one *E. coli*, and one *H. alvei* isolate in addition to the intrinsic resistance displayed for *Serratia* and *Morganella* spp. Upon repeat susceptibility testing, susceptibility for ertpapenem and meropenem were 18 and 52%, respectively. All but one of the isolates that were classified as unconfirmed-CRE were NCPE.

TABLE 1 | Distribution of carbapenemase and non-carbapenemase producing Enterobacterales.

	bla$_{KPC}$	*bla*$_{NDM}$	*bla*$_{OXA}$-type[b]	Other CPs[c]	CPE	NCPE	Unconfirmed
Number of isolates[a] *n* = 99; *n* (%)	30 (30)	5 (5)	2 (2)	5 (5)	41 (41)	58 (59)	17 (17)
Bacterial species; *n* (%)							
Klebsiella pneumoniae[a] (*n* = 41)	20 (49)	2 (5)	2 (5)	2 (5)	26 (63)	16 (39)	6 (15)
E. cloacae complex (*n* = 11)	1 (9)	1 (9)	0	0	2 (18)	9 (82)	1 (9)
Escherichia coli (*n* = 28)	5 (18)	2 (7)	0	0	7 (25)	22 (79)	7 (25)
Klebsiella aerogenes (*n* = 7)	2 (29)	0	0	0	2 (29)	5 (71)	1 (14)
Citrobacter freundii (*n* = 3)	1 (33)	0	0	0	1 (33)	2 (67)	0
Serratia marcescens[a] (*n* = 3)	1 (33)	0	0	3 (100)	3 (100)	0	0
Morganella morganii (*n* = 2)	0	0	0	0	0	2 (100)	1 (50)
Klebsiella michiganensis (*n* = 2)	0	0	0	0	0	2 (100)	1 (50)
Hafnia alvei (*n* = 1)	0	0	0	0	0	1 (100)	0

CP, carbapenemases; KPC, Klebsiella pneumoniae carbapenemase; NCPE, non-carbapenemase producing carbapenem-resistant Enterobacterales, NDM, New Delhi metallo-β-lactamase; OXA, oxacillinase.
[a]*A patient may carry more than one type of CPE.*
[b]*OXA-type includes OXA-48 and OXA-232.*
[c]*Other CP types include IMI-like, IMP-like, SME and VIM.*

The distribution of resistance determinants among CPE, NCPE and unconfirmed strains are summarized in **Supplementary Tables 2**, **3**. *K. pneumoniae* strains carried more *bla*$_{KPC}$, *bla*$_{CTX-M}$-variants, *bla*$_{SHV}$-variants, *fosA*-like, and *oqxA/B* genes compared with other Enterobacterales including *E. coli*. On the other hand, more *bla*$_{pAmpC}$ were prevalent in other Enterobacterales than those in *K. pneumoniae* strains. Notably, IncHI2-ST1 plasmids harboring the *mcr-9* gene was identified in 3 isolates (*E. cloacae* and *K. michiganensis*). Interestingly, these 3 isolates did not display *in-vitro* polymyxin non-susceptibility.

Among *K. pneumoniae* isolates, 62% were CPE with KPC-2 as the most prevalent enzyme. The two most common clonal groups (CG) of *K. pneumoniae* were CG258 [14 (33%) of 42] and CG307 [14 (33%) of 42]. CG258 represented 43% of the carbapenemase-producing *K. pneumoniae* (**Figure 1**). Of the 14 CG258 *K pneumoniae*, 79% were carbapenemase-producing, harboring primarily *bla*$_{KPC-2}$ (80%) or *bla*$_{KPC-3}$ (20%). The other predominant clonal group was CG307, which demonstrated a lower rate of carbapenemase-producing isolates [5 (36%)] as compared to CG258. In CG307 strains, *bla*$_{KPC-2}$ was the most common carbapenemase. One blood isolate from a patient who recently traveled to India harbored *bla*$_{OXA-232}$. In addition, all CG307 carried *bla*$_{CTX-M}$, a common group of extended-spectrum β-lactamases in this clonal group.

In *E. coli*, diverse genetic lineages were observed (**Figure 2**). No carbapenemase genes were present in 75% of *E. coli*. Notably, while ST131 has been reported as a common clonal type in this species, it was not identified in this diverse set. The carbapenemases in *E. coli* were *bla*$_{KPC-2}$ (n = 4), *bla*$_{KPC-3}$ (n = 1), and *bla*$_{NDM}$ (n = 2). Interestingly, two isolates containing *bla*$_{NDM-5}$ were identified in two independent cases of patients from Mexico. Both cases were ambulatory with minimal exposures to the healthcare system in Mexico prior to hospitalization. One case had a bronchial alveolar lavage performed from which an acid-fast bacilli stain was positive and identified as *Mycobacterium tuberculosis*-complex by polymerase chain reaction (PCR) in addition to *E. coli*. The second case was from a urine culture with polymicrobial growth that detected *E. coli*. In both cases, detection of *bla*$_{NDM}$ was made initially using the Cepheid Xpert® Carba-R PCR test.

Enterobacter cloacae isolates were the third most frequent group of CRE. We observed diverse genetic heterogeneity consistent with *E. cloacae* complex. No carbapenemase genes were present in most isolates [9 (82%) of 11]. All but one isolate harbored *bla*$_{AmpC}$ (*bla*$_{CMY}$ and *bla*$_{ACT/MIR}$) genes. Among these, six were found to have mutations in *ompF*. Among the two isolates with carbapenemase genes, one contained *bla*$_{KPC-3}$ and one harbored *bla*$_{NDM-5}$. The *bla*$_{NDM-5}$ was identified from a case of a patient with recent travel from Guatemala and no other health care exposures or significant comorbidities. No molecular tests were used at time of clinical care; susceptibility patterns displayed resistance to carbapenems and ceftazidime-avibactam but susceptible to tigecycline.

TABLE 2 | Distribution of carbapenemase and non-carbapenemase producing Enterobacterales and source of infection.

	bla$_{KPC}$	*bla*$_{NDM}$	*bla*$_{OXA}$-type	Other CPs	NCPE	*P (CPE vs. NCPE)*
Urine	14	1	1	0	12	0.0557
Wound swab	1	0	0	0	0	∼
Bone and soft tissues	4	1	0	0	6	0.773
Blood	1	0	1	0	6	0.325
Hepatobiliary specimen	1	0	0	0	6	0.123
Abdominal specimen	3	0	0	0	3	0.659
Respiratory specimens	3	0	0	0	12	0.044

[a]*Source of infection was not provided identified in 17 isolates.*

Epidemiology and Clinical Features

Phylogenetic reconstructions of *K. pneumoniae* and *E. coli* displaying types of NCPE and CPE demonstrated clonal heterogeneity. Three potential transmission clusters involving unique subjects were identified: one KPC-positive *K. pneumoniae*

TABLE 3 | Antimicrobial susceptibilities.

Name	All (%)	CPE (%)	NCPE (%)
AMI	91	91	91
AZT	11	9	12
CIP	19	16	21
DOX	54	69	41
ERT	18	2	28
FEP	26	19	41
FOT	11	16	9
GEN	51	44	53
MEM	52	22	69
MIN	63	81	52
TAZ	17	16	17
TGC	97	97	97
TOB	43	41	45
TZP	27	22	29

CPE, carbapenemase-producing Enterobacterales (n = 38), NCPE, non-carbapenemase-producing Enterobacterales (n = 58); AMI, Amikacin; AZT, Aztreonam, CIP, Ciprofloxacin; DOX, Doxyclicline; ERT, Ertapenem; FEP, Cefepime; FOT, Cefotetan; GEN, Gentamicin; MEM, Meropenem; MIN, Minocycline; TAZ, Ceftazidime; TGC, Tigecycline; TOB, Tobramycin; TZP, Piperacillin-tazobactam.

cluster (involving patients K702013 and K852013). Both isolates were *K. pneumoniae* ST307 [replicons IncFII(K)]. Another potential cluster of NCPE among three unique patients admitted on the same hospital floor was identified among *K. pneumoniae* ST307 (K2312011, K2542011, and K2342012).

Of the 53 individual patients with clinical data, 51% of patients were male and the median age was 55 years (interquartile range, 45–74 years; **Table 4**). Of the patients, 27 had acute care hospitalization and 26 had an emergency department visit in the prior 90 days. The majority of patients had antimicrobial use 90 days prior to culture. Approximately 47% (25/53) required admission to the intensive care unit. The all-cause 30-day mortality rate was 21% (11/53). At 30 days after culture date, 6% of patients were still hospitalized, 32% (17/53) were readmitted and were with lack of clinical response, 9% patients experienced relapse of their infection, 21% (11/53) remained on anti-CRE therapies, and 26% had developed worsening renal function. There were no differences in demographic or chronic conditions between patients with NCPE vs. CPE. Patients with NCPE (62%) were more likely to have had a prior ED visits as compared with patients with CPE ($p = 0.011$). The days of therapy (DOT) prior to culture was significantly higher among NCPE vs. CPE (795 vs. 242; $p \leq 0.001$). There were no differences between these categories as compared to unconfirmed CREs. We found no significant difference between mortality (29% vs. 17%, CPE vs NCPE; $p = 0.142$).

DISCUSSION

Carbapenem-resistant Enterobacterales presents a critical problem worldwide due to rapidly rising resistance rates and poor outcomes. The majority of efforts against CRE have focused on the epidemiology and outcomes in the context

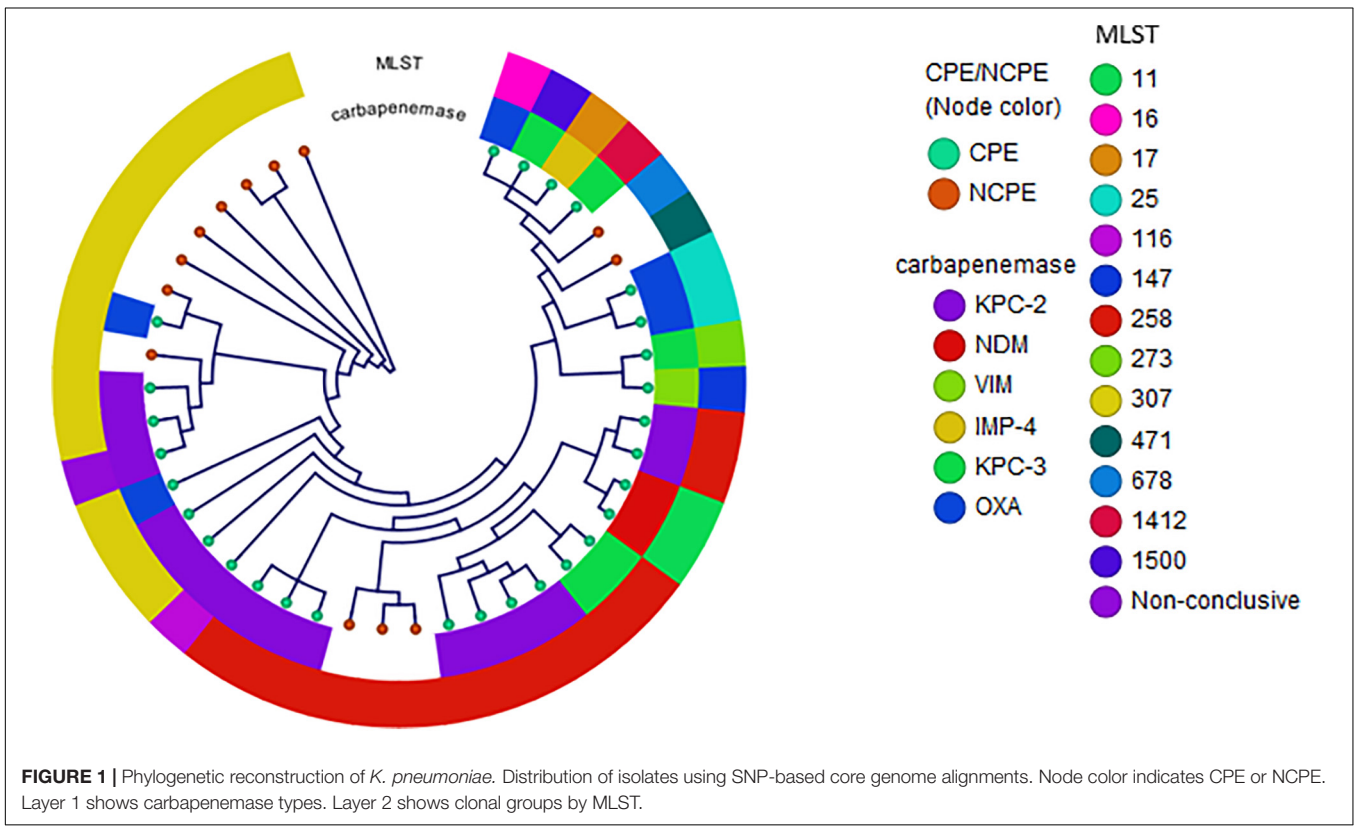


FIGURE 1 | Phylogenetic reconstruction of *K. pneumoniae.* Distribution of isolates using SNP-based core genome alignments. Node color indicates CPE or NCPE. Layer 1 shows carbapenemase types. Layer 2 shows clonal groups by MLST.

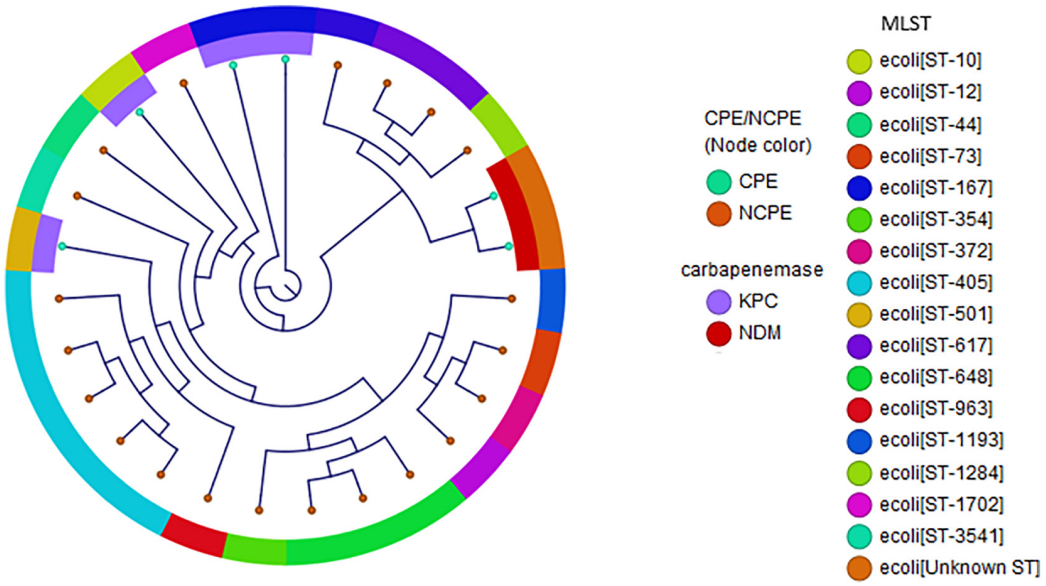


FIGURE 2 | Phylogenetic reconstruction of *E. coli.* Distribution of isolates using SNP-based core genome alignments. Node color indicates CPE or NCPE. Layer 1 shows carbapenemase types. Layer 2 shows clonal groups by MLST.

of CPE. However, in this cohort from South Texas, 48% of Enterobacterales that met the CDC definitions for CRE did not carry carbapenemase genes, and 17 were not carbapenem resistant upon repeat susceptibility testing. This has important implications for both clinical management as well as for epidemiologic reasons.

A recent multistate surveillance study estimated CRE rates of 2.93 per 100,000, with NCPE representing over 50% of the isolates and non-carbapenemase producing *K. pneumoniae* (NCPK) making up 58.6% of the total number of species (Duin et al., 2020). The current study found local rates of NCPK to be 27.6% of total NCPE. DOT were significantly higher in the NCPE group, indicating recent antimicrobial exposure as a contributing factor for the development of NCPE CRE. Furthermore, patients with NCPE were also associated with higher number of visits to the ED.

We report a high proportion of NCPE with varying mechanisms. In conjunction with porin loss or mutations, production of bla_{ESBL} and bla_{AmpC} have been shown to confer carbapenem-resistance, particularly in NCPK strains (Guh et al., 2015). This work validates this as bla_{KPC}, $bla_{\mathrm{CTX-M}}$-variants, and bla_{SHV}-variants were more abundant in the NCPK strains while bla_{pAmpCs} were more prevalent in other Enterobacterales (*E. coli* and *E. cloacae*). This work further confirms that both porin alterations and ESBL/AMPC production must be related responses to antimicrobial therapy in these strains. In NCPK, ertapenem and meropenem therapy failure have been shown to result from OmpK36 absence, with mutations playing a theoretical role (Hamzaoui et al., 2018). Moreover, *K. pneumoniae* outer membrane porin composition shifts during carbapenem therapy, specifically, the more restrictive OmpK36 porin seems to be favored over the wider OmpK35 porin in carbapenem-resistant strains (Findlay et al., 2012). The structure of OmpK36 is consistent with other Enterobacterales outer membrane proteins, consisting of 16 antiparallel β-barrel strands connected by short and long loops at the periplasmic and extracellular sides, respectively. However, wildtype OmpK36 extracellular loop 3 (L3) folds into the channel, partially obstructing the channel (Hasdemir et al., 2004). A solved crystal structure of OmpK36 from a meropenem-resistant strain has been shown to have a Gly115-Asp116 insertion within L3, stabilized by a salt bridge, further constricting meropenem's uptake compared to wildtype OmpK36 (Dutzler et al., 1999). Additionally, a premature OmpK36 stop codon at K103stop has been previously characterized in this phenotype, eliminating the critical extracellular loops and resulting in a null porin (Wong et al., 2019). *K. pneumoniae* strains which produce Ompk37 in an Ompk35/36-null membrane, have been shown to significantly select against β-lactam entry while maintaining nutrient uptake (Ruiz et al., 2012). OmpK37 achieves this selectivity by way of an inward-folded extracellular L3 loaded with fully or partially charged residues and a strategically oriented Y118 residue, shielding the channel further (**Supplementary Figure 1**; Ruiz et al., 2012). Consistent among β-lactam-resistant strains, our strains maintained the Y118 residue and had mutations within L3 and the OmpK37 channel which made for a more hydrophobic environment. However, the single polar-aromatic to basic nucleotide mutation at the intracellular side of the channel, Y311H, that has been observed in a small number of meropenem-resistant strains previously but was not discovered in our strains (Wong et al., 2019).

Molecular analysis of this clinical collection uncovered several emerging trends in this region. We note that since 2013, CG307 was associated with as many CRE infections as those of the

TABLE 4 | Clinical features of carbapenemase- and non-carbapenemase producing Enterobacterales.

	CPE (*n* = 17)	NCPE (*n* = 36)	*P*
Age (mean, SD)	49 (16)	51(11)	0.421
Male sex	10 (59)	17 (47)	0.212
Travel History	4 (23)	12 (33)	0.314
Race/Ethnicity			
White	12 (71)	28 (78)	0.570
Black	2 (12)	1 (3)	0.493
Hispanic	3 (18)	5 (14)	0.148
Comorbidities			
Charlson Comorbidity Score (mean, SD)	3.1 (0.6)	3.9 (0.8)	0.262
Myocardial infarct	0	0	–
Congestive heart failure	2 (12)	1 (3)	0.493
Peripheral vascular disease	5 (29)	17 (47)	0.352
Cerebrovascular disease	0	2 (6)	–
Dementia	0	0	–
Chronic pulmonary disease	2 (12)	2 (6)	0.809
Connective tissue disease	0	1 (3)	–
Ulcer disease	2 (12)	6 (17)	0.956
Mild liver disease	2 (12)	2 (6)	0.809
Diabetes (without complications)	3 (18)	6 (17)	0.761
Diabetes with end organ damage	1 (6)	6 (17)	0.517
Hemiplegia	0	4 (11)	–
Moderate or severe renal disease	6 (35)	11 (31)	0.976
Solid tumor (non metastatic)	1 (6)	5 (14)	0.693
Leukemia	1 (6)	0	–
Lymphoma, Multiple myeloma	0	0	–
Moderate or severe liver disease	2 (12)	1 (3)	0.493
Metastatic solid tumor	0	1 (3)	–
AIDS	0	0	–
HSCT – autologous	0	1 (3)	–
HSCT – allogenic	0	0	–
Hemodialysis	0	0	–
Setting prior to admission			
Home	12 (71)	25 (69)	0.932
Nursing Home/Long-term chronic care	1 (6)	7 (19)	0.198
Hospital Transfer	2 (12)	1 (3)	0.108
Long-term acute care	0	0	–
Transfer from different region/country	0	0	–
Hospice	0	0	–
Prior hospitalization or ED visits past 90 days			
Prior Hospitalization ($\geq$1)	6 (35)	21 (58)	0.146
Prior ED visits (>1)	4 (23)	22 (61)	0.011
Prior Antibiotics in last 90 days			
Fluoroquinolones	4 (23)	10 (28)	
3rd/4th Gen Cephalosporins	1 (6)	6 (17)	
1st Gen Cephalosporins	2 (12)	5 (14)	
Carbapenems	1 (6)	7 (19.4)	
Penicillins	0	9 (25)	
Macrolides	1 (6)	5 (14)	
Vancomycin	1 (6)	4 (11)	
Overall days of therapy	242	795	<0.001

CPE, carbapenemase-producing Enterobacterales; NCPE, non-carbapenemase-producing Enterobacterales; AIDS, acquired immunodeficiency syndrome; HCST, hematopoietic stem cell transplantation; ED, emergency department.

previous common clone, CG258. CG307 strains varied markedly in their gene content with carriage of $bla_{\text{CTX-M}}$ and bla_{KPC} although were more likely to be NCPE compared to CG258. This is consistent with trends from the Houston region, supporting the regional introduction and spread of this novel high-risk clone (Duin et al., 2020). Moreover, we identified two cases of *K. michiganensis*, a relatively new identified pathogen that is commonly mis-identified by clinical laboratories as *Klebsiella oxytoca* (Pagès et al., 2008). An isolate in this study was found to harbor the plasmid-borne gene *mcr-9* associated with polymyxin resistance. This is the first study to describe *mcr-9* harboring *K. michiganensis* in TX, United States.

We identified five strains expressing bla_{NDM} that conferred broad resistance to nearly all β-lactams. Of these, two isolates harboring bla_{NDM} were discovered in patients with recent travel history from Mexico. Recent studies have described outbreaks of $bla_{\text{NDM-1}}$ harboring nosocomial strains of *E. cloacae* and *K. pneumoniae* in Mexico City and other border cities (Founou et al., 2018). Global spread of bla_{NDM}-positive plasmids is concerning as these strains are associated with extremely reduced antimicrobial treatment options.

While the high proportion of NCPE was surprising, one explanation of this trend might be driven by the updated CDC definition of CRE in 2015. Other reports demonstrated an increase in "CRE" with the application of these definitions. In addition, we describe unconfirmed cases that met the CDC-defined definition (Duin et al., 2020). This subgroup may reflect some errors and differences in susceptibility testing methods. Furthermore, the expansion of the definition to include non-susceptible ertapenem (MIC breakpoint of 0.5 μg/mL) and inconsistencies between state and CDC definitions may have further contributed to discordant results. These could have implications for clinical microbiology laboratories. For example, all isolates that were intermediate resistant to ertapenem only while susceptible to all other carbapenems did not have detectable carbapenemases by the Cepheid Xpert® Carba-R PCR test. For this reason, our institutions do not routinely conduct these rapid molecular tests if non-susceptible to ertapenem only. The direct clinical impact of these various definitions, however, was not within the scope of this study. We found no significant difference between mortality between CPE vs NCPE. These findings are consistent with a recent prospective study found that clinical outcomes were not significantly different regardless of infection with CPE, NCPE, or unconfirmed CRE, with an overall mortality of 24% (Duin et al., 2020). Prior to the era of newer β-lactam/β-lactamase inhibitor combination, mortality rates from CRE reported to be up to 75% (Borer et al., 2009; van Duin et al., 2018; Wunderink et al., 2018; Alcántar-Curiel et al., 2019; Motsch et al., 2020).

There are several limitations to be noted in this study. First, information on clinical characteristics and outcomes could not be completely acquired for all unique patients because of the limitations of a retrospective study and migrating electronic health records. Second, the data herein is representative of two urban hospitals in South Texas; therefore, the generalizability is uncertain. However, it confers the importance of local/regional surveillance and epidemiology that may directly impact

diagnostic and treatment practices. The variations in detection and antimicrobial susceptibility testing between institutions and with the research laboratory may have introduced varying results attributing to misattribution and overestimation of CRE. Carbapenems are very labile and are usually the first to fail in Vitek and disk diffusion. Therefore, procedures for automatic repeat testing with concurrent quality control and new reagents were performed to resolve discrepancies.

In summary, we demonstrate the wide heterogeneity of resistance mechanisms for CRE in this region with a predominance of NCPE. Identification and distinction of these mechanisms by rapid species identification and subsequent carbapenemase characterization would allow for optimal treatment and infection control efforts.

MATERIALS AND METHODS

Bacterial Isolates

Ninety-nine CRE clinical isolates from unique patients at Methodist Hospital System and University Health System in San Antonio, TX, United States between 2011 and 2019 were included. The sources of the isolates included urine, blood, body fluid, pus, tracheal aspirate, sputum, wound swabs/tissue, central venous and arterial line tips. CRE was defined using the CDC definition (resistant to imipenem, meropenem, doripenem, or ertapenem OR documentation that the isolate possesses a carbapenemase) (Chea et al., 2015). Carbapenem non-susceptibility was defined based on CLSI breakpoints: meropenem MIC $\geq$ 2 μg/mL or ertapenem MIC $\geq$ 1 μg/mL (CLSI, 2018).

Carbapenem-resistant Enterobacterales with carbapenemase genes detected were termed CPE; those without carbapenemase genes were termed NCPE. We evaluated another category, termed unconfirmed-CRE, for CDC-defined CRE that although were identified as carbapenem-resistant by local laboratories, were found to be susceptible or intermediate to all tested carbapenems upon repeat testing.

Antimicrobial Susceptibility Testing

Isolates tested non-susceptible by Vitek2® (bioMérieux, Inc., NC, United States) or Etest® (bioMérieux, Inc., NC, United States) per CLSI breakpoints at the time of isolation from clinical specimens were included (i.e., ertapenem MIC $\geq$ 1 μg/mL or meropenem, imipenem, doripenem MICs of $\geq$2 mg/L). *In vitro* susceptibility for 20 antimicrobials (**Table 1**) was repeated with Thermo Scientific Sensititre™ GNX2F plate (Thermo Fisher Scientific, OH, United States) using broth microdilution method according to CLSI recommendations (M100-S28, 2018). Likewise, susceptibility was interpreted per CLSI except for colistin and tigecycline. CLSI epidemiological cutoff of $\geq$4 mg/L, which is also consistent with EUCAST resistant breakpoint of $>$2 mg/L, defined non-wild type for colistin and FDA breakpoints were used for tigecycline (i.e., Susceptible $\leq$2 mg/L, Intermediate 4 mg/L, Resistant $\geq$8 mg/L). *Pseudomonas aeruginosa* ATCC 27853 was used as a quality control organism.

Clinical Data Collection

Medical records were reviewed to collect clinical and epidemiologic data, including sample date, location of culture collection, specimen source, patient demographics, travel history, health care exposures, underlying medical conditions, indwelling devices and antimicrobial therapy exposures 90 days prior to culture, infection types, clinical manifestations, and outcomes. This study was reviewed and approved by the UT Health San Antonio Institutional Review Board.

DNA Sequencing and Analyses

Whole genome sequencing (WGS) was conducted on all 99 isolates using a NextSeq 500 sequencing instrument (Illumina Inc., San Diego, CA, United States) with 150-base paired-end reads (UT Health San Antonio, San Antonio, TX, United States). De-novo assembly and variant analyses were conducted using CLC Genomics Workbench 20.1 (Qiagen, Redwood City, CA, United States). The MLST database was used to identify the sequence types (STs) of the study isolates using WGS data. More detailed methods are provided in the **Supplementary Material**. Short-read data have been deposited in the NCBI BioProject (PRJNA688166).

Statistical Analyses

The student *t*-test or the nonparametric Wilcoxon rank-sum test was used for continuous variables based on distribution. The χ^2 or Fisher exact test was used to compare categorical variables. A two-sided *P*-value of less than 0.05 was considered statistically significant. All analyses were completed with SPSS 22 (IBMCorp).

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the NCBI repository, accession number (PRJNA688166).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Texas Health San Antonio. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

GL: conceptualization. GL, CB, and WS: analyses and draft the manuscript. CB, SA, GG, RB, CC, and JS: data extraction and experimentation. GL, CB, WS, SD, and GG: data interpretation. SD, GG, and RB: provide edits. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported in part by Merck (GL). The funders had no role in the design of the study or the analysis and

interpretation of the results. Data was generated in the Genome Sequencing Facility which is supported by UT Health San Antonio, NIH-NCI P30 CA054174 (Cancer Center at UT Health San Antonio), NIH Shared Instrument grant 1S10OD021805-01 (S10 grant), and CPRIT Core Facility Award (RP160732).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.623574/full#supplementary-material

REFERENCES

Alcántar-Curiel, M. D., Fernández-Vázquez, J. L., Toledano-Tableros, J. E., Gayosso-Vázquez, C., Jarillo-Quijada, M. D., López-Álvarez, M. D. R., et al. (2019). Emergence of IncFIA plasmid-carrying blaNDM-1 among *Klebsiella pneumoniae* and Enterobactercloacae isolates in a tertiary referral hospital in Mexico. *Microb. Drug Resist.* 25, 830–838. doi: 10.1089/mdr.2018.0306

Borer, A., Saidel-Odes, L., Riesenberg, K., Eskira, S., Peled, N., Nativ, R., et al. (2009). Attributable mortality rate for carbapenem-resistant *Klebsiella pneumoniae* bacteremia. *Infect. Control Hosp. Epidemiol.* 30, 972–976. doi: 10.1086/605922

Centers for Disease Control and Prevention (U.S.) (2019). *Antibiotic Resistance Threats in the United States, 2019*. Atlanta, GA: Centers for Disease Control and Prevention (U.S.). doi: 10.15620/cdc:82532

Chea, N., Bulens, S. N., Kongphet-Tran, T., Lynfield, R., Shaw, K. M., Vagnone, P. S., et al. (2015). Improved phenotype-based definition for identifying carbapenemase producers among carbapenem-resistant Enterobacteriaceae. *Emerg. Infect. Dis.* 21, 1611–1616. doi: 10.3201/eid2109.150198

CLSI (2018). Performance Standards for Antimicrobial Susceptibility Testing—Twenty-Eighth Edition: M100. CLSI. Wayne

Duin, D., van Arias, C. A., Komarow, L., Chen, L., Hanson, B. M., Weston, G., et al. (2020). Molecular and clinical epidemiology of carbapenem-resistant Enterobacterales in the USA (CRACKLE-2): a prospective cohort study. *Lancet Infect. Dis.* 20, 731–741. doi: 10.1016/S1473-3099(19)30755-8

Dutzler, R., Rummel, G., Albertí, S., Hernández-Allés, S., Phale, P., Rosenbusch, J., et al. (1999). Crystal structure and functional characterization of OmpK36, the osmoporin of *Klebsiella pneumoniae*. *Structure* 7, 425–434. doi: 10.1016/S0969-2126(99)80055-0

Findlay, J., Hamouda, A., Dancer, S. J., and Amyes, S. G. B. (2012). Rapid acquisition of decreased carbapenem susceptibility in a strain of *Klebsiella pneumoniae* arising during meropenem therapy. *Clin. Microbiol. Infect.* 18, 140–146. doi: 10.1111/j.1469-0691.2011.03515.x

Founou, R. C., Founou, L. L., Allam, M., Ismail, A., and Essack, S. Y. (2018). Genomic characterisation of *Klebsiella michiganensis* co-producing OXA-181 and NDM-1 carbapenemases isolated from a cancer patient in uMgungundlovu District, KwaZulu-Natal Province, South Africa. *S. Afr. Med. J.* 109, 7–8. doi: 10.7196/SAMJ.2018.v109i1.13696

Guh, A. Y., Bulens, S. N., Mu, Y., Jacob, J. T., Reno, J., Scott, J., et al. (2015). Epidemiology of carbapenem-resistant Enterobacteriaceae in 7 US communities, 2012-2013. *JAMA* 314, 1479–1487. doi: 10.1001/jama.2015.12480

Hamzaoui, Z., Ocampo-Sosa, A., Fernandez Martinez, M., Landolsi, S., Ferjani, S., Maamar, E., et al. (2018). Role of association of OmpK35 and OmpK36 alteration and blaESBL and/or blaAmpC genes in conferring carbapenem resistance among non-carbapenemase-producing *Klebsiella pneumoniae*. *Int. J. Antimicrob. Agents* 52, 898–905. doi: 10.1016/j.ijantimicag.2018.03.020

Hasdemir, U. O., Chevalier, J., Nordmann, P., and Pagès, J.-M. (2004). Detection and prevalence of active drug efflux mechanism in various multidrug-resistant *Klebsiella pneumoniae* strains from Turkey. *J. Clin. Microbiol.* 42, 2701–2706. doi: 10.1128/JCM.42.6.2701-2706.2004

Motsch, J., Oliveira, C., Stus, V., Köksal, I., Lyulko, O., Boucher, H. W., et al. (2020). RESTORE-IMI 1: a multicenter, randomized, double-blind trial comparing efficacy and safety of imipenem/relebactam vs colistin plus imipenem in patients with imipenem-nonsusceptible bacterial infections. *Clin. Infect. Dis.* 70, 1799–1808. doi: 10.1093/cid/ciz530

Pagès, J., James, C. E., and Winterhalter, M. (2008). The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria. *Nat. Rev. Microbiol.* 6, 893–903. doi: 10.1038/nrmicro1994

Ruiz, E., Ocampo-Sosa, A. A., Rezusta, A., Revillo, M. J., Román, E., Torres, C., et al. (2012). Acquisition of carbapenem resistance in multiresistant *Klebsiella pneumoniae* strains harbouring blaCTX-M-15, qnrS1 and aac(6′)-Ib-cr genes. *J. Med. Microbiol.* 61, 672–677. doi: 10.1099/jmm.0.038083-0

van Duin, D., Lok, J. J., Earley, M., Cober, E., Richter, S. S., Perez, F., et al. (2018). Colistin versus ceftazidime-avibactam in the treatment of infections due to carbapenem-resistant Enterobacteriaceae. *Clin. Infect. Dis.* 66, 163–171. doi: 10.1093/cid/cix783

WHO (2020). *Global Priority List Of Antibiotic-Resistant Bacteria To Guide Research, Discovery, And Development Of New Antibiotics WHO*. Available online at: http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/ (accessed September 30, 2020)

Wong, J. L. C., Romano, M., Kerry, L. E., Kwong, H.-S., Low, W.-W., Brett, S. J., et al. (2019). OmpK36-mediated carbapenem resistance attenuates ST258 *Klebsiella pneumoniae* in vivo. *Nat. Commun.* 10, 1–10. doi: 10.1038/s41467-019-11756-y

Wunderink, R. G., Giamarellos-Bourboulis, E. J., Rahav, G., Mathers, A. J., Bassetti, M., Vazquez, J., et al. (2018). Effect and safety of meropenem-vaborbactam versus best-available therapy in patients with carbapenem-resistant Enterobacteriaceae infections: the TANGO II randomized clinical trial. *Infect. Dis. Ther.* 7, 439–455. doi: 10.1007/s40121-018-0214-1

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

Copyright © 2021 Black, So, Dallas, Gawrys, Benavides, Aguilar, Chen, Shurko and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership