frontiers | Research Topics

Integration of NGS in clinical and public health microbiology workflows: Applications, compliance, quality considerations

Edited by
Varvara K. Kozyreva, Shangxin Yang, Ruth Evangeline Timme,
Peera Hemarajata and Heather A. Carleton

Published in
Frontiers in Public Health
Frontiers in Cellular and Infection Microbiology
Frontiers in Microbiology

ISSN 1664-8714
ISBN 978-2-8325-4419-8
DOI 10.3389/978-2-8325-4419-8

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

Integration of NGS in clinical and public health microbiology workflows: Applications, compliance, quality considerations

Topic editors

Varvara K. Kozyreva — California Department of Public Health, United States
Shangxin Yang — University of California, Los Angeles, United States
Ruth Evangeline Timme — US Food and Drug Administration, United States
Peera Hemarajata — Association of Public Health Laboratories, United States
Heather A. Carleton — Centers for Disease Control and Prevention (CDC), United States

Citation

Kozyreva, V. K., Yang, S., Timme, R. E., Hemarajata, P., Carleton, H. A., eds. (2024). *Integration of NGS in clinical and public health microbiology workflows: Applications, compliance, quality considerations*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4419-8

Table of contents

frontiers | Frontiers in Public Health

TYPE Editorial
PUBLISHED 23 January 2024
DOI 10.3389/fpubh.2024.1357098

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Marc Jean Struelens,
Université Libre de Bruxelles, Belgium

*CORRESPONDENCE
Shangxin Yang
✉ shangxinyang@mednet.ucla.edu

RECEIVED 17 December 2023
ACCEPTED 15 January 2024
PUBLISHED 23 January 2024

CITATION
Yang S, Kozyreva VK, Timme RE and Hemarajata P (2024) Editorial: Integration of NGS in clinical and public health microbiology workflows: applications, compliance, quality considerations.
Front. Public Health 12:1357098.
doi: 10.3389/fpubh.2024.1357098

Editorial: Integration of NGS in clinical and public health microbiology workflows: applications, compliance, quality considerations

Shangxin Yang[1]*, Varvara K. Kozyreva[2], Ruth E. Timme[3] and Peera Hemarajata[4]

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [2]Microbial Diseases Laboratory, California Department of Public Health, Richmond, CA, United States, [3]Center for Food Safety and Applied Nutrition, United States Food and Drug Administration, College Park, MD, United States, [4]Association of Public Health Laboratories, Silver Spring, MD, United States

KEYWORDS

next-generation sequencing (NGS), clinical microbiology, public health microbiology, whole-genome sequencing (WGS), metagenomics, outbreak investigation, genomic surveillance, infectious diseases

Editorial on the Research Topic
Integration of NGS in clinical and public health microbiology workflows: applications, compliance, quality considerations

Since its invention in the 90s, next-generation sequencing (NGS) has played an instrumental role in pushing our understanding of human microbiome and microbial genomics to a whole new level (1, 2). In the past decade, NGS has also been widely adopted in public health and food safety laboratories and became the primary method for microbial surveillance and outbreak investigation (3, 4). This trend extends to the clinical laboratories, where NGS has been a powerful tool for hospital outbreak investigation and institutional-level pathogen surveillance to aid infection prevention programs (5, 6). Soon after, the exploration of NGS's diagnostic utility for infectious diseases gained tremendous momentum, with both whole-genome sequencing (WGS) based and metagenomics (mNGS) based tests showing dramatic improvements in the detection and identification of pathogens that otherwise couldn't be detected or accurately identified, thus solving unmet clinical needs (7–9). Despite the great promise and indisputable values, integration of NGS is a challenging endeavor not only for each individual laboratory but also for our entire field. The technical complexity, the lack of guidelines and standards, and the extraordinary resources required are some of the most remarkable obstacles (10, 11). In this Research Topic, integration and utilization of NGS in clinical and public health microbiology was described in great detail, encompassing wet-lab techniques, bioinformatics, logistics, outbreak investigation, genomic surveillance, and patient diagnostics.

One highlight of this Research Topic is the use of fungal WGS for genomic surveillance, which historically had been less established compared to bacterial genomic surveillance. Using WGS, Michael et al. demonstrated that the infamous "Black Molds" epidemic during the delta wave of the COVID-19 pandemic in India was caused by multiple fungal species, predominantly *Rhizopus delemar*. Even among this species there was vast genetic diversity indicating no common sources nor a particular strain. Most COVID-19 patients who suffered mucormycosis had diabetes which is a known risk factor for both COVID-19 infection and invasive mucormycosis. These findings were significant as they demystified the role of Mucorales and its relationship with COVID-19. In another important study, Gorzalski et al. utilized 3 different bioinformatics tools to analyze the WGS data of over 200 *Candida auris* isolates from an ongoing large outbreak in Nevada and elucidated the inferred transmission networks based on shared SNP analysis. This study provided essential genomic epidemiological data to help understand the dynamics of this large outbreak which brought unprecedented challenges to the hospitals in the affected areas even to date. This study also highlighted the importance of implementing real-time genomic surveillance of *Candida auris* to help slow down its transmission.

The bioinformatics workflow TheiaEuk described in the manuscript by Ambrosio et al. was designed to fully utilize the benefit of a cloud computing platform. Their work on *C. auris* genomic epidemiological analysis involved a large dataset, which demonstrated that cloud computing is perhaps the only truly scalable and sustainable solution for bioinformatic analyses.

Analyzing fungal WGS data is challenging, yet the interpretation of fungal phylogenetic results can be equally hard, as demonstrated in another outbreak investigation by Fan et al.. In this study, *Cyberlindnera fabianii*, an unusual yeast was recovered from the urine culture of 3 patients from the same ward, prompting a suspicion for nosocomial infections. SNP analysis revealed that two of the *C. fabianii* isolates had 192 SNPs difference while the third was over 26,000 SNPs apart. The main conundrum was how to interpret this 192 SNPs distance; the authors did a literature review showing that the genetic difference of yeast isolates with epidemiological link could range widely from <50 SNPs to >1000 SNPs, depending on the genome size of the species and length of the outbreaks. Given the similar size of *C. fabianii* compared to *C. auris*, 192 SNPs could still be interpreted as likely having a common source. Ultimately, the only way to solve this type of interpretation dilemma is to sequence many more fungal pathogens and pair it with extensive collection of epidemiological information on the potential transmission chains, which will expand the fungal genome database and our knowledge base of epidemiology of fungal nosocomial outbreaks and fungal evolution during infection and colonization.

The improvement in result turnaround time and increasing accessibility of sequencing technologies, even in limited-resources circumstances, allows researchers to find innovative ways to diagnose and improve the quality of care for high-consequence endemic diseases such as tuberculosis. In this Research Topic, we showcased seminal work by two different groups related to the use of targeted NGS (tNGS) for the diagnosis and prediction of antimicrobial resistance directly from primary specimens. tNGS approach for TB resistance prediction is clearly favorable for clinical application due to its ability to generate actionable results with a rapid turnaround time, as opposed to whole-genome sequencing, which is a great surveillance tool, but requires pure culture, hence, leading to delays in obtaining the results. The work by Murphy et al. described in detail a clinically validated, state-of-the-art approach to using tNGS coupled with real-time long read sequencing technology and customized bioinformatic pipeline to examine genes and mutations in *Mycobacterium tuberculosis* to predict resistance to antimicrobials, allowing clinicians to choose the most appropriate treatment for each patient weeks before WGS results were available. The work performed by de Araujo et al. further emphasized the potential feasibility of utilizing tNGS to enhance clinical and surveillance efforts to combat drug-resistant *M. tuberculosis* by outlining an innovative programmatic framework that incorporated *M. tuberculosis* tNGS in low-resource regions where NGS had not previously been available.

NGS has been widely used in surveillance of foodborne pathogens and healthcare associated pathogens, especially ones with resistant mechanisms of public health concern, like carbapenem-resistant organisms (CRO). Yet the utilization of genomic data for identification of outbreak sources and efficient communication of genomic results to the epidemiologists still could be improved to make genomic epidemiology truly actionable. Gali et al. highlights an application of automated, NGS cluster analysis tool at NCBI Pathogen Detection, which provides public health investigators current, pre-computed clustering data commonly used for the investigation of foodborne outbreaks. The Virginia Division of Consolidated Laboratory Services (DCLS) laboratory has extended this application to detect and identify the sources outbreaks of CRO, specifically involving *Acinetobacter baumannii, Enterobacter cloacae, Morganella morganii, Klebsiella pneumoniae, Escherichia coli*, and *Proteus mirabilis*.

The proper identification of pathogens is a cornerstone of successful surveillance as well as clinical diagnosis. However, even with such common pathogens as enteric bacteria, the scarcity of well-curated reference datasets impedes clinical validations of identification tools as well as the development of new bioinformatics solutions. The paper presented by Lindsey et al. described the development of Reference Genome Dataset for benchmarking of enteric genomic identification using Average Nucleotide Identity (ANI) algorithm. The manuscript also provides a nice example of clinical validation of a bioinformatics tool, including determination of genome coverage limits for successful ANI identification.

The mNGS technology has revolutionized pathogen detection. Historically, many fastidious or endemic pathogens have been under detected due to lack of effective diagnostic tools. As an agnostic "one-stop test," mNGS is shown by Chang et al. to be particularly powerful in diagnosing a case of *Leishmania donovani* visceral leishmaniasis, a rare infectious disease unexpectedly found in an infant with acute lymphoblastic leukemia. The timely diagnosis led to successful treatment, demonstrating the value of mNGS. Another difficult-to-detect microorganism, *Legionella pneumonia,* was successfully identified with mNGS in a clinical case presented by Du et al. in which Legionnaires' disease coincided with rhabdomyolysis and acute kidney injury, a.k.a. Legionella Triad, a rare and deadly syndrome requiring timely diagnosis and treatment. The patient finally improved, and the authors advocate for the implementation of mNGS for the early diagnosis of severe cases of Legionnaires' disease in resource limited areas.

Question of clinical relevance of mNGS findings remains topical and matter of relative and absolute abundance of detected species and its importance in making clinical interpretations regarding the role of the detected pathogens as causative agents of the disease are widely discussed. Unlike aforementioned studies describing implementation of hypothesis-free mNGS approach, Almas et al. showcase the use of a hybridization capture-based sequencing method for the diagnosis of urinary tract infections (UTIs) by combining broad detection range benefits of NGS technology and precision of targeted approach for focusing data generation on clinically relevant information. The ability of the bioinformatic platform presented by the authors to provide quantitative results is particularly attractive for clinical microbiology applications.

Implementation of NGS is not without a set of unique challenges, which include but are not limited to requirement of expensive equipment and reagent, availability of skilled scientists to perform the wet lab part, access to high-performance computing platforms and well-trained bioinformaticians, and an effective way to validate and communicate results to clinicians and epidemiologists. Tartanian et al. astutely described not only their trials and tribulations in implementing SARS-CoV-2 sequencing within their health system, but also many aspects of their efforts that were incredibly successful. With the sheer volume of COVID-19 samples at the height of the pandemic, most, if not all entities performing SARS-CoV-2 sequencing had to find creative ways to increase the throughput of their sequencing efforts. One of the approaches taken by institutions with financial support was to implement automation throughout the NGS workflow in efforts to increase the throughput. Socea et al. described their success story, albeit not without some challenges, in implementing automated library preparation to eliminate one of the potential major bottlenecks in the NGS process. These stories of overcoming challenges to establish next-generation sequencing (NGS) capabilities provide valuable practical insights for those facing similar odds.

NGS, as a tool for pathogen genomic surveillance, requires tight coordination at multiple levels to ensure the NGS data are of sufficient quality and associated contextual data meet the requirements for public health action, both globally and locally. These coordinating efforts often include both public and private databases, increasing the complexity of data management in submitting and extracting data for public health action. Wadford et al. and the State of California established the California SARS-CoV-2 Whole Genome Sequencing (WGS) Initiative, or “California COVIDNet.” This cross-sector collaboration implemented large-scale genomic surveillance of SARS-CoV-2 across California to monitor the spread of variants within the state, to detect new and emerging variants, and to characterize outbreaks in various settings, including congregate facilities and workplaces. The framework and computational infrastructure developed for COVID-19 can be extended now for pathogen genomic surveillance of other infectious diseases.

Four decades ago, polymerase chain reaction (PCR) was invented and now it has become a primary diagnostic and screening tool for infectious diseases; NGS undoubtedly is following the same trajectory. Many efforts and innovations are still required to lower the cost and hurdles for the integration of NGS in clinical and public health microbiology, yet a bright future lays ahead. The advancements in automation, bioinformatics and database curation, and better consensus and guidelines for implementation of NGS assays in regulated environments for clinical testing will accelerate the widespread adoption of NGS and strengthen our capabilities for fighting the infectious diseases with ever-changing landscape.

Author contributions

SY: Conceptualization, Writing—original draft, Writing—review & editing. VK: Conceptualization, Writing—original draft, Writing—review & editing. RT: Conceptualization, Writing—original draft, Writing—review & editing. PH: Conceptualization, Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

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

Publisher’s note

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

References

1. Weinstock GM. Genomic approaches to studying the human microbiota. *Nature.* (2012) 489:250–6. doi: 10.1038/nature11553

2. Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* (2016) 17:333–51. doi: 10.1038/nrg.2016.49

3. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin Infect Dis.* (2016) 63:380–6. doi: 10.1093/cid/ciw242

4. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol.* (2019) 17:533–45. doi: 10.1038/s41579-019-0214-5

5. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK, et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci Transl Med.* (2012) 4:148ra116. doi: 10.1126/scitranslmed.3004129

6. Yang S, Hemarajata P, Hindler J, Li F, Adisetiyo H, Aldrovandi G, et al. Evolution and transmission of carbapenem-resistant klebsiella pneumoniae expressing the blaOXA-232 gene during an institutional outbreak associated with endoscopic retrograde cholangiopancreatography. *Clin Infect Dis.* (2017) 64:894–901. doi: 10.1093/cid/ciw876

7. Thoendel MJ, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Identification of prosthetic joint infection pathogens using a shotgun metagenomics approach. *Clin Infect Dis.* (2018) 67:1333–8. doi: 10.1093/cid/ciy303

8. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet.* (2019) 20:341–55. doi: 10.1038/s41576-019-0113-7

9. Salem-Bango Z, Price TK, Chan JL, Chandrasekaran S, Garner OB, Yang S. Fungal whole-genome sequencing for species identification: from test development to clinical utilization. *J Fungi (Basel).* (2023) 9. doi: 10.3390/jof9020183

10. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio.* (2015) 6:e01888–01815. doi: 10.1128/mBio.01888-15

11. Price TK, Realegeno S, Mirasol R, Tsan A, Chandrasekaran S, Garner OB, et al. Validation, implementation, and clinical utility of whole genome sequence-based bacterial identification in the clinical microbiology laboratory. *J Mol Diagn.* (2021) 23:1468–77. doi: 10.1016/j.jmoldx.2021.07.020

frontiers | Frontiers in Cellular and Infection Microbiology

TYPE Original Research
PUBLISHED 07 March 2023
DOI 10.3389/fcimb.2023.1130645

Check for updates

OPEN ACCESS

EDITED BY
Shangxin Yang,
University of California, Los Angeles,
United States

REVIEWED BY
Marie Desnos-Ollivier,
Institut Pasteur, France
Andres Ceballos-Garzon,
Université Clermont Auvergne, France

*CORRESPONDENCE
Li Gu
guli2013227@foxmail.com

[†]These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to
Clinical Microbiology,
a section of the journal
Frontiers in Cellular and
Infection Microbiology

RECEIVED 23 December 2022
ACCEPTED 21 February 2023
PUBLISHED 07 March 2023

CITATION
Fan X, Dai R-C, Kudinha T and Gu L (2023)
A pseudo-outbreak of *Cyberlindnera fabianii* funguria: Implication from whole genome sequencing assay.
Front. Cell. Infect. Microbiol. 13:1130645.
doi: 10.3389/fcimb.2023.1130645

A pseudo-outbreak of *Cyberlindnera fabianii* funguria: Implication from whole genome sequencing assay

Xin Fan[1†], Rong-Chen Dai[2†], Timothy Kudinha[3,4] and Li Gu[1*]

[1]Department of Infectious Diseases and Clinical Microbiology, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing, China, [2]School of Public Health, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China, [3]School of Dentistry and Medical Sciences, Charles Sturt University, Leeds Parade, Oranges, NSW, Australia, [4]NSW Health Pathology, Regional and Rural, Orange hospital, Orange, NSW, Australia

Background: Although the yeast *Cyberlindnera fabianii* (*C. fabianii*) has been rarely reported in human infections, nosocomial outbreaks caused by this organism have been documented. Here we report a pseudo-outbreak of *C. fabianii* in a urology department of a Chinese hospital over a two-week period.

Methods: Three patients were admitted to the urology department of a tertiary teaching hospital in Beijing, China, from Nov to Dec 2018, for different medical intervention demands. During the period Nov 28 to Dec 5, funguria occurred in these three patients, and two of them had positive urine cultures multiple times. Sequencing of rDNA internal transcribed spacer (ITS) region and MALDI-TOF MS were applied for strain identification. Further, sequencing of rDNA non-transcribed spacer (NTS) region and whole genome sequencing approaches were used for outbreak investigation purpose.

Results: All the cultured yeast strains were identified as *C. fabianii* by sequencing of ITS region, and were 100% identical to the *C. fabianii* type strain CBS 5640T. However, the MALDI-TOF MS system failed to correctly identify this yeast pathogen. Moreover, isolates from these three clustered cases shared 99.91%-100% identical NTS region sequences, which could not rule out the possibility of an outbreak. However, whole genome sequencing results revealed that only two of the *C. fabianii* cases were genetically-related with a pairwise SNP of 192 nt, whilst the third case had over 26,000 SNPs on its genome, suggesting a different origin. Furthermore, the genomes of the first three case strains were phylogenetically even more diverged when compared to a *C. fabianii* strain identified from another patient, who was admitted to a general surgical department of the same hospital 7 months later. One of the first three patients eventually passed away due to poor general conditions, one was asymptomatic, and other clinically improved.

Conclusion: In conclusion, nosocomial outbreaks caused by emerging and uncommon fungal species are increasingly being reported, hence awareness must be raised. Genotyping with commonly used universal gene targets may have limited discriminatory power in tracing the sources of infection for these organisms, requiring use of whole genome sequencing to confirm outbreak events.

KEYWORDS

uncommon fungal pathogen, molecular typing, nosocomial transmission, *Cyberlindnera fabianii*, whole genome sequencing (WGS)

Introduction

Emerging fungal infections have become a global health concern in the past few decades due to their notable morbidity and mortality, especially among immunosuppressed patients admitted to intensive care units (ICUs), or undergoing invasive medical interventions (Pappas et al., 2018; Hoenigl et al., 2022; World Health Organization, 2022). Although *Candida albicans* remains the most predominant yeast pathogen, the incidence of uncommon yeast species causing human infections has increased enormously in recent years (Pappas et al., 2018; Chen et al., 2021). Uncommon yeast species often exhibit decreased susceptibility to commonly used antifungal agents, making them difficult to manage in clinical settings. Moreover, there are increasing incidences of nosocomial infections and outbreak events reported due to transmission of uncommon or emerging yeast species worldwide (Pappas et al., 2018; Chen et al., 2021). For instance, *Candida auris*, which was first described in 2009, has caused a number of outbreaks in different continents (Chow et al., 2018; Hoenigl et al., 2022; World Health Organization, 2022).

In the investigations and tracing of fungal nosocomial transmissions, molecular genotyping could provide essential genetic evidence. Hence, a wide variety of molecular typing assays have been evaluated and implemented in the study of outbreaks, including band-pattern-based DNA analysis like random amplified polymorphic DNA (RAPD) and pulsed field gel electrophoresis (PFGE), traditional DNA sequencing-based phylogenetic methods like single gene analysis, microsatellite analysis and multilocus sequence typing (MLST), protein spectrum analysis by matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS), and whole genome sequencing (WGS) techniques (Reiss et al., 1998; Pulcrano et al., 2012; Mikosz et al., 2014; Xiao et al., 2014; Litvintseva et al., 2015; Oliveira and Azevedo, 2022). Of these methods, WGS has become increasingly used due to its outstanding discriminatory power (Litvintseva et al., 2015; Bougnoux et al., 2018; Desnos-Ollivier et al., 2020; Oliveira and Azevedo, 2022).

In this study, we report on three clustered funguria cases caused by a rare fungal pathogen, *Cyberlindnera fabianii*, which occurred over a two-week period within the same urology department, which was initially considered as a nosocomial outbreak event. Using WGS, this event was finally confirmed as a pseudo-outbreak caused by *C. fabianii* from two diverged genetic lineages.

Material and methods

Ethics

This study was approved by the Human Research Ethics Committee of the Beijing Chaoyang Hospital (No. KE332). Written informed consent was obtained from all participants involved.

Routine isolation of the microorganisms and MALDI-TOF MS identification.

C. fabianii strains were isolated from urine samples of three patients (number 1-3); on four different occasions for patient number 1, only once for patient number 2, and on three different occasions for patient number 3 (Figure 1 and Table 1). After these three cases, no further *C. fabianii* cases were detected in the same hospital for over seven months, till a new *C. fabianii* isolate, cultured from an ascites sample of a patient admitted to general surgery department (recorded as patient number 4), was detected, and this isolate was used as a comparator.

Routine culture of specimens was carried out as per standard laboratory protocols. Generally, for urine samples, 1 μl of the sample was inoculated on Blood Agar media and incubated at 35 °C for 24 h. Thereafter, the number of colonies growing on the media plate was counted to ensure that they met the criterion for a urinary tract infection. A brief identification protocol revealed that the cultured isolates were yeast. Thus, Sabouraud glucose agar (SDA) was used to subculture these isolates for further identification testing. Attempts were made to identify the cultured yeasts by using a Vitek MS MALDI-TOF MS system (bioMérieux, Marcy l'Etoile, France, with IVD database version 2.1), following manufacturer's instructions. For each run, *Escherichia coli* strain ATCC 8739 was used to calibrate and control the method. Unfortunately, this system was unable to identify the yeast strains.

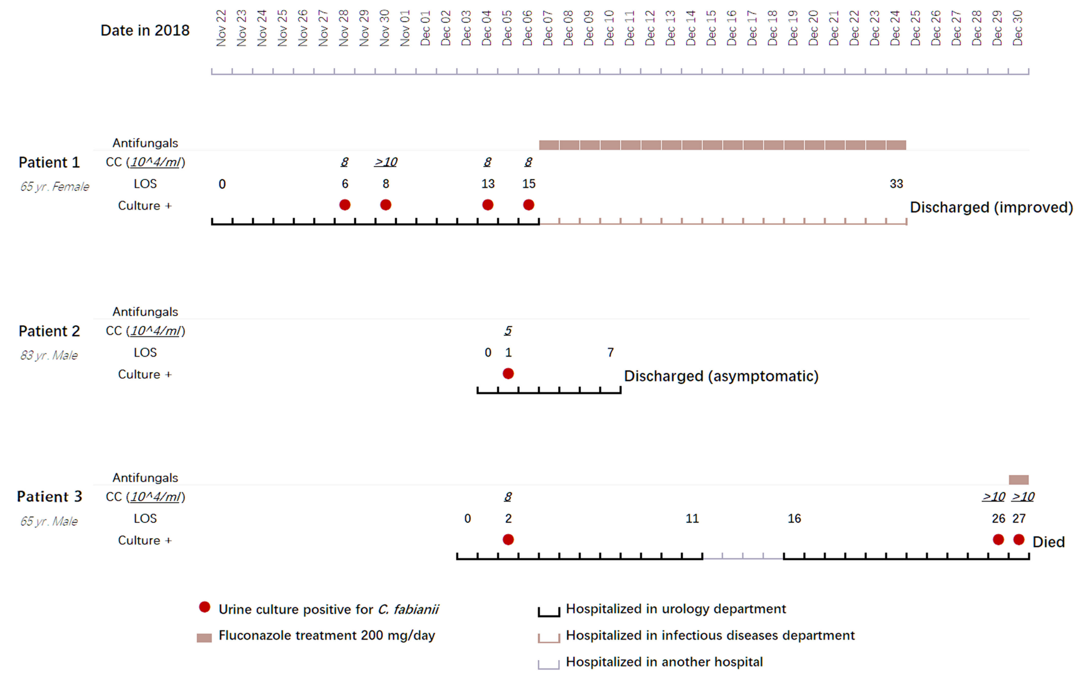


FIGURE 1
Clinical features, treatment regimens, and outcomes of three clustered cases with *Cyberlindnera fabianii* funguria. Abbreviations: CC, CFU (colony forming unit) count; LOS, length of stay; Culture +: culture positive for *C. fabianii*.

Molecular identification and phylogenetic analysis by rDNA gene spacer regions

As all the suspected yeast isolates could not be identified using the MALDI-TOF MS systems, sequencing of rDNA internal transcribed spacer (ITS) regions was carried out. Briefly, DNA extraction of the isolates was performed using a QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). The universal primer pair ITS1 and ITS4 was used for amplification and sequencing of the ITS region for each strain (Xiao et al., 2014), and a species identification was carried out by querying against the Westerdijk Fungal Biodiversity Institute's database using a web-based pairwise alignment tool (https://wi.knaw.nl/page/Pairwise_alignment).

Further, to investigate the potential relatedness of these cases, the first yeast isolate of each patient case was chosen for further testing, and the rDNA non-transcribed spacer region 1 (NTS-1) was amplified with a forward primer NTS1-F (5'-GGGATAAATCATTTGTATACGAC-3') and a reverse primer NTS1-R (5'-TTGCGGCCATATCCACAAGAAA-3') as described previously (Al-Sweih et al., 2019), and sequenced from both directions. A phylogenetic tree of NTS-1 sequences was generated by Mega X (version 10.2, https://www.megasoftware.net/) using neighbor-joining method with bootstrap value of 1000. NTS-1 sequences from *C. fabianii* type strain CBS 5640T, and *C. fabianii* reference genome strain JOY008, were also downloaded from GenBank and included in the analysis. In addition, NTS-1 sequence extracted from the genome of *Cyberlindnera jadinii* strain NBRC 0988 was selected as an outgroup.

Whole genome sequencing and analysis of *C. fabianii* strains

Whole-genome sequencing was performed on each of the first yeast strain from each of patients 1 to 4. Generally, a 350-bp DNA library was prepared using NEB Next Ultra DNA library prep kits (NEW ENGLAND BioLabs Inc., MA, USA), following the manufacturer's instructions. Library integrity was evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Sequencing was performed on an Illumina NovaSeq using PE150 in a commercial company (Novogene Co., Ltd., Beijing, China).

For genome analysis, the complete reference genome of *C. fabianii* strain JOY008 (GenBank accession no. GCA_022641835.1) was used for read mapping. Single-nucleotide polymorphism (SNP) analysis was carried out by Burrows-Wheeler Aligner (version 0.7.7), SAMtools (version 1.2), and Genome Analysis Toolkit (GATK) (v.3.3-0) per GATK Best Practices (Li and Durbin, 2009; Li et al., 2009; Mckenna et al, 2010). The filtered reads were compared to the reference genome by SAMtools to generate BAM files. Then, variants were marked by GATK MarkDuplicates for each sample, and single-sample GVCF files were created by GATK HaplotypeCaller with the option –emitRefConfidence GVCF. The GVCF files were aggregated by GATK CombineGVCFs tool. After that, the GVCF files were jointly genotyped with the GATK GenotypeGVCFs to produce a single VCF file containing variants data on every strain. Finally, the VCF file was selected using GATK SelectVariants with the option -select-type SNP and filtered using the following parameters: VariantFiltration, QD < 2.0,

TABLE 1 Clinical features of four patients with *Cyberlindnera fabianii* funguria and microbiology characteristics of the strains.

Patient	No. 1	No. 2	No. 3	No. 4
Patient features				
Reason for hospitalization	Fever and parastomal fistula	Postoperative follow-up of bladder cancer	Fever with backaches	Pancreatic cancer
Underlying disease	Bladder cancer	Bladder cancer	Bladder cancer, diabetes	No
Clinical status at time of first positive culture				
Fever	Yes	No	Yes	Yes
Immunosuppressive state	Yes	Yes	Yes	Yes
Neutropenia (<10^9/L)	No	No	No	No
High urine leukocytes	Yes	No	Yes	No
Prior antibacterial exposure	Yes	Yes	Yes	Yes
Prior antifungal exposure	No	No	No	No
Abdominal surgery	Yes	Yes	Yes	Yes
Indwelling urinary catheter	No	No	Yes	No
Parenteral nutrition	No	No	No	Yes
Yeast culture				
Department of hospitalization	Urology	Urology	Urology	General surgery
Samples positive for yeasts	Urine	Urine	Urine	Ascites fluid
Number of times isolated	4	1	3	1
Mixed bacteria culture positive	*Enterococcus faecium*	No	*Enterococcus faecium*	*Enterococcus faecium*, *Enterobacter cloacae*
Identification				
Lab no. of first strain	CYCFB01-1	CYCFB02-1	CYCFB03-1	CYCFB04-1
ITS sequencing	*C. fabianii*	*C. fabianii*	*C. fabianii*	*C. fabianii*
Identity with type strain	100%	100%	100%	100%
Vitek MS	No identification	No identification	No identification	No identification
Antifungal susceptibility (mg/L)				
Fluconazole	1	1	0.5	1
Voriconazole	0.015	0.015	0.015	0.03
Itraconazole	0.12	0.12	0.06	0.06
Posaconazole	0.12	0.12	0.12	0.12
Caspofungin	0.03	0.06	0.03	0.06
Micafungin	0.03	0.03	0.03	0.03
Anidulafungin	0.015	0.06	0.015	0.015
5- Flucytosine	0.12	0.06	0.12	0.06
Amphotericin B	0.5	0.5	0.25	0.25
Data availability				
ITS	OP904191	OP904192	OP904193	OP904194
NTS-1	OP912967	OP912968	OP9129689	OP91296870
WGS	SAMN32011978	SAMN32011979	SAMN320119810	SAMN32011981

ITS, rDNA internal transcribed spacer region; NTS-1, rDNA non-transcribed spacer region-1; WGS, whole genome sequencing.

ReadPosRankSum < –8.0, FS > 60.0, MQRankSum < –12.5, MQ < 40.0 and HaplotypeScore > 13.0.

Specifically, in this study, the term "pseudo-outbreak" was used to describe inappropriate artifactual clustering of real infections as an outbreak event, due to limitation of investigation tools.

Antifungal susceptibility testing

Minimum inhibitory concentrations (MICs) of all the isolates were determined by Sensititre YeastOne YO10 kits (Thermo Scientific, OH, USA) following the manufacturer's instructions, and with *Candida krusei* ATCC 6258 and *Candida parapsilosis* ATCC 22019, used as quality control strains.

Data availability

DNA sequences of rDNA ITS and NTS-1 regions for each of the first yeast strain isolated from each individual has been deposited into NCBI GenBank database (accession nos. OP904191-OP904194 for ITS region and OP912967-OP912970 for NTS-1 region). Their WGS reads data is also now available in National Microbiology Data Center (NMDC) database (Bioproject accession no. PRJNA907923).

Results

Patients

Information pertaining to each of the 4 patients included in this case study is summarized in Figure 1 and Table 1.

Patient 1 was a 65-year-old female admitted to the urology department of Beijing Chao-Yang hospital on Nov 22, 2018, due to presence of fever for two weeks and a parastomal fistula after ileal replacement due to bladder cancer in 2016. Upon admission, the patient had fever for over a week. On day 6 after admission, a yeast strain was isolated from her urine sample and the colony count (CC) was 8×10^4 CFU/ml. The same urine culture also grew *Enterococcus faecium* (5×10^4 CFU/ml) as a mixed culture with the yeast. The routine MALDI-TOF MS identification system failed to identify the yeast isolate. Her urine samples collected on days 8, 13 and 15 after admission also yielded yeasts (CC of 8 to >10×10^4 CFU/ml). Follow-up ITS sequencing assigned all the strains as *C. fabianii*. The patient was given fluconazole at 200 mg/day for 18 days after which her condition improved notably, and she was finally discharged from the hospital on day 33 of admission.

Patient 2 was an 83-year-old male admitted to the urology department of the hospital on Dec 04, 2018, for follow-up of bladder cancer electrosurgery performed eight and four months before his admission. On day 1 after admission, a urine sample was collected for routine screening, which was reported positive for yeasts with a CC of 5×10^4 CFU/ml. The yeast strain was identified as *C. fabianii* by ITS sequencing. This patient didn't present with any symptoms of infection, and hence antifungal therapy was not given. Later, he received a transurethral resection of bladder tumor on day 3, and was discharged on day 7 after admission.

Patient 3 was a 65-year-old male admitted to the urology department of the hospital on Dec 03, 2018, due to presence of high fever with backaches. Nine months before this admission, the patient hand undergone nephroureterectomy of the left kidney. He received nephrostomy on the right kidney immediately on day 1 after admission. His urine sample collected on day 2 was culture positive for a yeast (CC: 8×10^4 CFU/ml), which was identified as *C. fabianii* by ITS sequencing. However, no antifungal agents were prescribed for him and only a broad-spectrum antibiotic was given. On days 26 and 27, two urine samples were collected consecutively and both were positive for *C. fabianii* with a CC of > 10×10^4 CFU/ml. Of note, all his urine samples also grew *E. faecium* (>10×10^4 CFU/ml) as part of a mixed culture with the yeast. Though fluconazole therapy (200 mg/day) was initiated on day 27 after admission, the patient passed away on the same day.

Patient 4 was a 54-year-old female admitted to the general surgery department of the hospital on Jul 18, 2019, which was over seven months after the patient 1, 2 and 3 case clusters. She was hospitalized due to pancreatic cancer, and received radical pancreatoduodenectomy on day 12 after admission; later with pancreatic intestinal anastomotic fistula. The patient's ascites sample collected on day 20 was reported positive for *C. fabianii* and *E. faecium.* However, she didn't have any other culture positive results for fungi after that, nor received any antifungal treatment, and was discharged from the hospital on day 52.

C. fabianii identification

All the yeast strains could not be identified using the Vitek MS MALDI-TOF MS system IVD 2.0 database, nor were the isolates misidentified as something else (identification confidence values <60.0). This is not surprising as *C. fabianii* is not currently included in the system's spectrum database.

By sequencing of the ITS region, all the yeast strains from the four patients were unambiguously assigned to *C. fabianii*, with their respective ITS sequences 100% (602/602 bp) identical to that of *C. fabianii* type strain CBS 5640T.

Phylogenetic analysis by rDNA NTS-1 region

Since *C. fabianii* is a rare yeast species identified in human patients, and the fact that the clustered cases (patients 1 to 3) described here were identified within a two-week period from the same department, an investigation was carried out to assess the possibility of this being a nosocomial outbreak. Owing to the high sequence similarity of the ITS region among the strains, sequence analysis based on rDNA NTS-1 region was further carried out, which was assumed to have higher discriminatory power and has previously been used to confirm a *C. fabianii* outbreak in Kuwait (Al-Sweih et al., 2019).

Using *C. jadinii* (strain NBRC 0988) as an outgroup, the phylogenetic tree based on the NTS-1 region clustered all the *C. fabianii* isolates together (Figure 2), and inter-species variation between *C. jadinii* and *C. fabianii* in NTS-1 region was >43.5%. Amongst *C. fabianii* strains, some intra-species variation was observed (Figure 2). However, sequence variations amongst the strains from patients 1 to 3 was inconspicuous, as these strains exhibited the same sequence type, while the strain from patient 2 had only one SNP (identity 1179/1180, 99.92%). In contrast, strains from the three clustered cases were quite diverged from the strain from patient 4 which was isolated seven months later, which had an overall 7-bp insertions and 4 additional SNPs in its NTS-1 region (identity 99.07%).

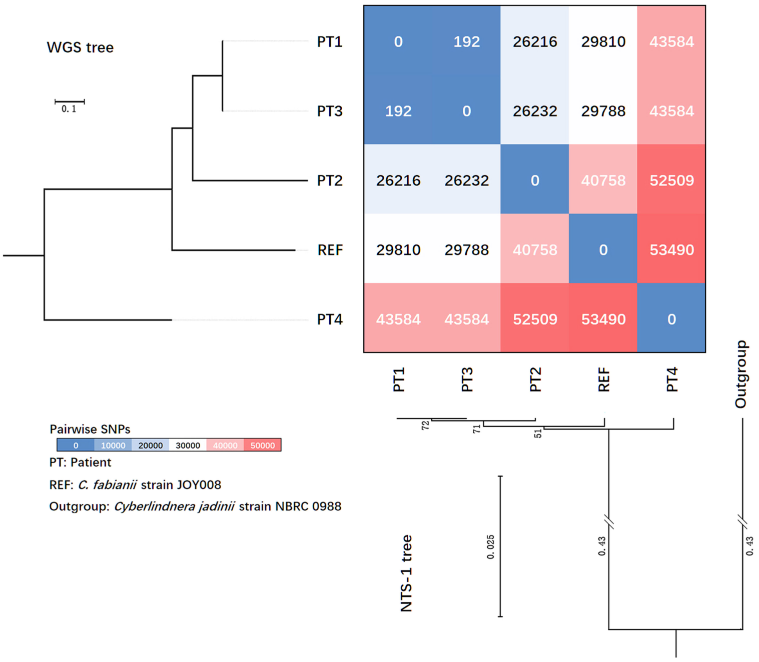
Whole genome sequencing results

Genome sequencing of yeast strains from patients 1 to 4 produced 2.6 to 3.7 Gb of clean data, and average depths of sequencing were all above 200×. The average genome size obtained was 12.94 Mb. Their genomes had an average GC content of 44.4% to 45.1%, with N50 of 13,739 bp to 202,514 bp. Comparative genomic analysis was performed for all strains. The pairwise differences between genome reference strain JOY008, which originated from a soil environment in USA, and our four patients' clinical strains, were 29,810-53,490 SNPs (Figure 2).

We carried out a review of previous outbreak reports caused by different yeast species, and the number of pairwise SNPs described varied from less than ten to several hundred (Table 2). The yeast strains from patient 1 and 3 had only 192 SNPs identified, suggesting that they were probably closely related (Figure 2). However, there were over 26,000 SNPs identified between the strain from patient 2 and strains from patients 1 and 3 (Figure 2), which was considered as a high-level genomic variation. These findings suggested that the *C. fabianii* strain from patient 2 was from a different origin. In addition, the yeast strain from patient 4 was even more diverged, with >43,000 of SNPs compared to all strains from patients 1 to 3, and the reference genome (Figure 2). Lastly, the phylogenetic tree constructed based on whole genome SNPs also support the same conclusion (Figure 2).

Antifungal susceptibilities

All the *C. fabianii* strains isolated in this study showed good susceptibility to all the nine antifungal agents tested (Table 1), with geometric minimum inhibitory concentration (GM MIC) of 0.84 mg/L to fluconazole, 0.02 mg/L to voriconazole, 0.08 mg/L to itraconazole, 0.12 mg/L to posaconazole, 0.04 mg/L to caspofungin, 0.03 mg/L to micafungin, 0.02 mg/L to anidulafungin, 0.08 mg/L to 5-flucytosine, and finally, 0.35 mg/L to amphotericin B. If using clinical breakpoints or epidemiological cut-off values of *C. albicans* as references, all these strains could be classified as susceptible or of wild-type phenotype to all antifungal agents tested.



	PT1	PT3	PT2	REF	PT4
PT1	0	192	26216	29810	43584
PT3	192	0	26232	29788	43584
PT2	26216	26232	0	40758	52509
REF	29810	29788	40758	0	53490
PT4	43584	43584	52509	53490	0

FIGURE 2
Phylogenetic trees generated based on rDNA non-transcribed spacer (NTS) region-1 sequences and whole genome sequencing (WGS) SNPs, and heatmaps revealing pairwise differences of SNPs amongst four patients' strains collected in this study.

TABLE 2 Review of outbreak events caused by yeast species that were characterized by whole genome sequencing in previous studies.

Species	Reference Genome size (Mb)	Country	Patient population	Ward	No. of cases with WGS	No. of SNPs within each event	Reference
Candida albicans	14.3	Spain	Neonate	ICU	2-11	134-769	(Guinea et al., 2021)
Candida parapsilosis	13	Spain	Neonate	ICU	2-4	49-241	(Guinea et al., 2021)
Candida auris	12.7	US	Adults	Not specified	26	2-50	(De St Maurice et al., 2022)
		India	Adults	Medical wards	2-2	≤7	(Yadav et al., 2021)
		UK	Adults	ICU, high dependency units, surgical admission ward	5-17	≤134	(Rhodes et al., 2018)
		UK	Adults	ICU, neurosciences wards	37	≤215	(Eyre et al., 2018)
		Colombia	Not specified	Not specified	5	≤40	(Escandon et al., 2019)
		USA	Not specified	Not specified	10	≤12	(Chow et al., 2018)
Dirkmeia churashimaensis	21	India	Neonate	ICU	6	1,621	(Chowdhary et al., 2020a)
Candida blankii	14.8	India	Neonate	ICU	6	≤277	(Chowdhary et al., 2020b)
Malassezia pachydermatis	8.2	USA	Neonate	ICU	5	≤14	(Chow et al., 2020)
Cyberlindnera fabianii	12.3	China	Adults	Urology department	2	192	This study

Discussion

C. fabianii, basionym *Hansenula fabianii*, homotypic synonyms *Candida fabianii*, *Lindnera fabianii* and *Pichia fabianii*, is an ascomycetous yeast that has a close relationship with human activities (Kato et al., 1997; Arastehfar et al., 2019; Van Rijswijck et al., 2019). This yeast species is commonly seen in fermented food products like alcohols (Arastehfar et al., 2019; Van Rijswijck et al, 2019), and has also been used for treatment of waste water with a long history (Kato et al., 1997). The species has now been assigned within the Wickerhamomycetaceae clade (Kidd et al., 2023). Within this clade, there are several other species that have been reported to cause human infections, such as *Wickerhamomyces anomalus* and *Cyberlindnera jadinii* (Treguier et al., 2018; Zhang et al., 2021).

Generally, detection of *C. fabianii* in clinical settings is rare. According to previous surveillance reports on human fungal diseases, the prevalence of *C. fabianii* is generally <0.1% (Pfaller et al., 2019; Xiao et al., 2020). However, this yeast species is also an opportunistic pathogen that can cause a broad-range of infections, including lethal fungemia (Al-Sweih et al., 2019; Arastehfar et al., 2019). A previous research suggests that *C. fabianii* only has low virulence attributes (Arastehfar et al., 2019), although Nouraei et al. observed that this fungal species was one of the uncommon yeasts with high-level production of hemolysin, phospholipase and proteinase (Nouraei et al., 2020). In addition, *C. fabianii* has been observed to have a strong capacity for biofilm formation, which may contribute to its persistence and resistance to antifungal therapies (Hamal et al., 2008; Nouraei et al., 2020).

Of note, several studies have revealed difficulties in the accurate identification of *C. fabianii* using conventional methods, which may influence precision clinical recognition and management of infections caused by this organism (Svobodova et al., 2016; Al-Sweih et al., 2019). MALDI-TOF MS has been reported as a powerful tool for identification of yeasts, but the system's identification capacity largely relies on the peptide mass fingerprint database that is incorporated into the system. Some of the MALDI-TOF MS systems, such as Biotyper (Bruker Daltonics, Germany, with IVD library version 8) and MicroIDSys (ASTA, Korea, with database version 1.23.2), have demonstrated capacity to accurately identify *C. fabianii* strains (Park et al., 2019; Teke et al., 2021). In contrast to this, *C. fabianii* is still absent from the Vitek MS IVD database (up to version 3.2), hence this system failed to identify any of *C. fabianii* isolates in this study. Similar findings were also reported by Teke et al. (Teke et al., 2021).

Although nosocomial transmission of fungal pathogens is less frequently encountered in clinical practice compared to bacterial pathogens, fungal outbreaks are more common than publicly appreciated, and are mostly associated with medical products or contamination of the hospital environment (Kanamori et al., 2015; Litvintseva et al., 2015; Magill et al., 2018). For instance, *Candida parapsilosis*, one of the most prevalent human pathogenic yeast species, is well-known for causing catheter-related bloodstream infections. There have been a large number of nosocomial outbreaks caused by *Candida parapsilosis* worldwide, including several recently reported cases caused by fluconazole-resistant clones that raised more public health concerns (Arastehfar et al., 2020; Zhang et al., 2020; Thomaz et al., 2022). Moreover, reports of outbreaks caused by unusual fungal pathogens, such as the recently emerged *C. auris*, are increasing (Litvintseva et al., 2015; Chow et al., 2018). Of note, during the COVID-19 pandemic period, fungal outbreaks caused higher medical burdens to healthcare facilities and patients (Hoenigl et al., 2022; Thomaz et al., 2022), and hence are beginning to receive more attention.

Of note, a recent outbreak of *C. fabianii* in Kuwait was described by Al-Sweih et al., which involved a total of 10 fungemia cases in neonates (Al-Sweih et al., 2019). Furthermore, previous reviews on *C. fabianii* cases have demonstrated that the elderly population is the second most vulnerable population after neonates overall, with funguria being the first to second commonest clinical symptom (Al-Sweih et al., 2019; Arastehfar et al., 2019; Park et al., 2019). This agrees with our three-clustered *C. fabianii* funguria cases which all occurred in elderly patients, and with funguria as the common clinical symptom, though not every patient had symptomatic urinary tract infection.

Published literature have emphasized that presence of indwelling urinary catheter is the most important risk factor and transmission route for nosocomial urinary tract infections, especially when catheter care quality is poor. However, a variety of additional risk factors have also been described, including female gender, increased age, diabetes, bladder instrumentation, urinary outflow obstruction, amongst others. (Pearson-Stuttard et al., 2016; Mody et al., 2017; Odabasi and Mert, 2020). Of the four patients described in this study, only one carried an indwelling urinary catheter, and all of them had undergone abdominal surgeries prior to the onset of funguria. Besides, three of the four patients had *E. faecium* detected concurrently with *C. fabianii* in the same urine sample. *Enterococcus* species, including *E. faecium*, are well-known ubiquitous inhabitants of the human gut microbiota and could lead to urinary tract infections (Magruder et al., 2019). Moreover, *C. fabianii* has also been identified in the human intestinal microbiota (Zhai et al., 2020), and previously Mathy et al. hypothesized that translocation of *C. fabianii* from the gut was responsible for a ventriculoperitoneal shunt case (Mathy et al., 2020). Therefore, it is possible that *C. fabianii* funguria cases identified in our study may have resulted from gut microbiota translocations, and abdominal surgeries might serve as triggers or risk factors.

As widely-acknowledged, application of ITS sequencing could allow accurate identification of yeast species but with insufficient discriminatory power for intra-species typing (Stielow et al., 2015; Al-Sweih et al., 2019). Al-Sweih et al. applied sequencing of NTS-1 regions, a gene locus that is considered to have a higher discriminatory power, in *C. fabianii* outbreak investigation, and found that all outbreak strains in Kuwait shared 100% identical NTS-1 sequences (Al-Sweih et al., 2019). In comparison, we found a single SNP within NTS-1 region on patient 2's strain versus strains from patients 1 and 3 in this study. However, further solid evidence was still needed to rule out the possibility of a potential outbreak.

To address concerns on readiness and limitations in discriminatory power of molecular typing methods in outbreak investigations, WGS has been recommended as a valuable alternative (Litvintseva et al., 2015; Bougnoux et al., 2018; Desnos-Ollivier et al., 2020). In this study, SNP-based analysis based on results acquired from WGS data clearly suggested that the genome of patient 2's strain was quite divergent amongst the three clustered cases, which indicated a pseudo-outbreak event. Of note, the phrase "pseudo-outbreaks" could refer to either clustering of false infections, or artifactual clustering of real infections (Wallace et al., 1998). Clustering of false infections was more widely-noted, which may be associated with e.g. medical device or clinical laboratory contaminations (Kirby et al., 2017; Abdolrasouli et al., 2021). However, as indicated in our study, artifactual misinterpretation of "outbreaks" due to limitation of investigation methodologies (such as inadequate discriminatory power of molecular typing assays) should also be avoided.

Although WGS has made significant contributions in epidemiological studies, some limitations still remain. One major issue, as noted in outbreak investigations of all microbes including bacteria and fungi, is lack of consensus for data interpretation. Specifically, setting-up pairwise SNP-based cut-off values for assigning transmission events is still cumbersome, which has limited the wide utility of WGS in epidemiological studies (Coll et al., 2020; Guinea et al., 2021). In review of previous reports for outbreaks caused by yeast species that were characterized by WGS, it can be seen that the number of pairwise SNPs described in different studies of diverged species varied significantly, from less than ten to over hundreds. In the present study, genomic evidence clearly supported that patient 2's *C. fabianii* strain was from a different origin, compared to others (with >26,000 SNPs compared to strains from patients 1 and 3). However, the 192 pairwise-SNP between strains from patient 1 and 3 may suggest that these two patients could have acquired the yeasts from a common source in the same ward but through different routes, rather than a direct transmission between the 2 patients, in which case the number of SNPs would be expected to be much less. But the hypothesis needs additional evaluation in a larger population and with more cases.

Due to the possibility of nosocomial transmission of this yeast in the described ward, surveillance infection control cultures were obtained to screen for *C. fabianii* in the department's environment and amongst related health-care staff, but no *C. fabianii* was detected. Additional infection control strategies implemented further included enhancing environmental cleaning and hand hygiene practices, as well as providing education of fungal nosocomial infections to all healthcare staff.

One limitation of the study is that, antifungal susceptibility testing was not carried out using the standard broth microdilution methods, though YeastOne has proved equally efficient with good correlation in testing of yeasts (Cuenca-Estrella et al., 2010). Furthermore, with the limited number of cases studied, our base-line understanding for intra-species variation of *C. fabianii* genomes was still limited. Interpretation of any outbreak events shouldn't simply rely on WGS result alone. It warrants a comprehensive analysis of different aspects of the cases, including patients' clinical characteristics and epidemiological data, as well as the pathogens' phenotypic and molecular characteristics.

In conclusion, as there are increasing reports of nosocomial outbreaks caused by emerging and uncommon fungal species, increased awareness of these rare organisms is warranted in public health. Conventional genotyping methods may have limited discriminatory power in investigating outbreaks due to these rare organisms; WGS has proven to be a good typing method for supporting investigation of such rare outbreak events.

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: Genbank: OP904191-OP904194 for ITS region, OP912967-OP912970 for NTS-1 region. WGS reads data can be found in NCBI database under Bioproject accession no. PRJNA907923.

Ethics statement

The studies involving human participants were reviewed and approved by Human Research Ethics Committee of the Beijing Chaoyang Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

XF, R-CD and LG conceived the work. XF, and R-CD performed the experiments and data analysis. XF, TK, and LG drafting the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Beijing Hospitals Authority Youth Program (grant no. QML20190301) and Reform and Development Program of Beijing Institute of Respiratory Medicine (XF and LG).

Conflict of interest

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

Publisher's note

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

References

Abdolrasouli, A., Gibani, M. M., De Groot, T., Borman, A. M., Hoffman, P., Azadian, B. S., et al. (2021). A pseudo-outbreak of rhinocladiella similis in a bronchoscopy unit of a tertiary care teaching hospital in London, united kingdom. *Mycoses* 64, 394–404. doi: 10.1111/myc.13227

Al-Sweih, N., Ahmad, S., Khan, S., Joseph, L., Asadzadeh, M., and Khan, Z. (2019). Cyberlindnera fabianii fungaemia outbreak in preterm neonates in Kuwait and literature review. *Mycoses* 62, 51–61. doi: 10.1111/myc.12846

Arastehfar, A., Daneshnia, F., Hilmioglu-Polat, S., Fang, W., Yasar, M., Polat, F., et al. (2020). First report of candidemia clonal outbreak caused by emerging fluconazole-resistant candida parapsilosis isolates harboring Y132F and/or Y132F +K143R in Turkey. *Antimicrob. Agents Chemother.* 64, e01001–20. doi: 10.1128/AAC.01001-20

Arastehfar, A., Fang, W., Al-Hatmi, A. M. S., Afsarian, M. H., Daneshnia, F., Bakhtiari, M., et al. (2019). Unequivocal identification of an underestimated opportunistic yeast species, cyberlindnera fabianii, and its close relatives using a dual-function PCR and literature review of published cases. *Med. Mycol* 57, 833–840. doi: 10.1093/mmy/myy148

Bougnoux, M. E., Brun, S., and Zahar, J. R. (2018). Healthcare-associated fungal outbreaks: New and uncommon species, new molecular tools for investigation and prevention. *Antimicrob. Resist. Infect. Control* 7, 45. doi: 10.1186/s13756-018-0338-9

Chen, S. C., Perfect, J., Colombo, A. L., Cornely, O. A., Groll, A. H., Seidel, D., et al. (2021). Global guideline for the diagnosis and management of rare yeast infections: an initiative of the ECMM in cooperation with ISHAM and ASM. *Lancet Infect. Dis.* 21, e375–e386. doi: 10.1016/S1473-3099(21)00203-6

Chow, N. A., Chinn, R., Pong, A., Schultz, K., Kim, J., Gade, L., et al. (2020). Use of whole-genome sequencing to detect an outbreak of malassezia pachydermatis infection and colonization in a neonatal intensive care unit-california-2016. *Infect. Control Hosp Epidemiol.* 41, 851–853. doi: 10.1017/ice.2020.73

Chow, N. A., Gade, L., Tsay, S. V., Forsberg, K., Greenko, J. A., Southwick, K. L., et al. (2018). Multiple introductions and subsequent transmission of multidrug-resistant candida auris in the USA: a molecular epidemiological survey. *Lancet Infect. Dis.* 18, 1377–1384. doi: 10.1016/S1473-3099(18)30597-8

Chowdhary, A., Sharada, K., Singh, P. K., Bhagwani, D. K., Kumar, N., De Groot, T., et al. (2020a). Outbreak of dirkmeia churashimaensis fungemia in a neonatal intensive care unit, India. *Emerg. Infect. Dis.* 26, 764–768. doi: 10.3201/eid2604.190847

Chowdhary, A., Stielow, J. B., Upadhyaya, G., Singh, P. K., Singh, A., and Meis, J. F. (2020b). Candida blankii: an emerging yeast in an outbreak of fungaemia in neonates in Delhi, India. *Clin. Microbiol. Infect.* 26 648, e645–648e648. doi: 10.1016/j.cmi.2020.01.001

Coll, F., Raven, K. E., Knight, G. M., Blane, B., Harrison, E. M., Leek, D., et al. (2020). Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant staphylococcus aureus: a genomic epidemiology analysis. *Lancet Microbe* 1, e328–e335. doi: 10.1016/S2666-5247(20)30149-X

Cuenca-Estrella, M., Gomez-Lopez, A., Alastruey-Izquierdo, A., Bernal-Martinez, L., Cuesta, I., Buitrago, M. J., et al. (2010). Comparison of the vitek 2 antifungal susceptibility system with the clinical and laboratory standards institute (CLSI) and European committee on antimicrobial susceptibility testing (EUCAST) broth microdilution reference methods and with the sensititre YeastOne and etest techniques for *in vitro* detection of antifungal resistance in yeast isolates. *J. Clin. Microbiol.* 48, 1782–1786. doi: 10.1128/JCM.02316-09

Desnos-Ollivier, M., Maufrais, C., Pihet, M., Aznar, C., Dromer, F., and French Mycoses Study, G. (2020). Epidemiological investigation for grouped cases of trichosporon asahii using whole genome and IGS1 sequencing. *Mycoses* 63, 942–951. doi: 10.1111/myc.13126

De St Maurice, A., Parti, U., Anikst, V. E., Harper, T., Mirasol, R., Dayo, A. J., et al. (2022). Clinical, microbiological, and genomic characteristics of clade-III candida auris colonization and infection in southern california-2022. *Infect. Control Hosp Epidemiol.* 1-9. doi: 10.1017/ice.2022.204

Escandon, P., Chow, N. A., Caceres, D. H., Gade, L., Berkow, E. L., Armstrong, P., et al. (2019). Molecular epidemiology of candida auris in Colombia reveals a highly related, countrywide colonization with regional patterns in amphotericin B resistance. *Clin. Infect. Dis.* 68, 15–21. doi: 10.1093/cid/ciy411

Eyre, D. W., Sheppard, A. E., Madder, H., Moir, I., Moroney, R., Quan, T. P., et al. (2018). And its control in an intensive care setting. *N Engl. J. Med.* 379, 1322–1331. doi: 10.1056/NEJMoa1714373

Guinea, J., Mezquita, S., Gomez, A., Padilla, B., Zamora, E., Sanchez-Luna, M., et al. (2021). Whole genome sequencing confirms candida albicans and candida parapsilosis microsatellite sporadic and persistent clones causing outbreaks of candidemia in neonates. *Med. Mycol* 60. doi: 10.1093/mmy/myab068

Hamal, P., Ostransky, J., Dendis, M., Horvath, R., Ruzicka, F., Buchta, V., et al. (2008). A case of endocarditis caused by the yeast pichia fabianii with biofilm production and developed *in vitro* resistance to azoles in the course of antifungal treatment. *Med. Mycol* 46, 601–605. doi: 10.1080/13693780802078180

Hoenigl, M., Seidel, D., Sprute, R., Cunha, C., Oliverio, M., Goldman, G. H., et al. (2022). COVID-19-associated fungal infections. *Nat. Microbiol.* 7, 1127–1140. doi: 10.1038/s41564-022-01172-2

Kanamori, H., Rutala, W. A., Sickbert-Bennett, E. E., and Weber, D. J. (2015). Review of fungal outbreaks and infection prevention in healthcare settings during construction and renovation. *Clin. Infect. Dis.* 61, 433–444. doi: 10.1093/cid/civ297

Kato, M., Iefuji, H., Miyake, K., and Iimura, Y. (1997). Transformation system for a wastewater treatment yeast, hansenula fabianii J640: isolation of the orotidine-5'-phosphate decarboxylase gene (URA3) and uracil auxotrophic mutants. *Appl. Microbiol. Biotechnol.* 48, 621–625. doi: 10.1007/s002530051105

Kidd, S. E., Abdolrasouli, A., and Hagen, F. (2023). Fungal nomenclature: Managing change is the name of the game. *Open Forum Infect. Dis.* 10, ofac559. doi: 10.1093/ofid/ofac559

Kirby, J. E., Branch-Elliman, W., Lasalvia, M. T., Longhi, L., Mackechnie, M., Urman, G., et al. (2017). Investigation of a candida guilliermondii pseudo-outbreak reveals a novel source of laboratory contamination. *J. Clin. Microbiol.* 55, 1080–1089. doi: 10.1128/JCM.02336-16

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

Litvintseva, A. P., Brandt, M. E., Mody, R. K., and Lockhart, S. R. (2015). Investigating fungal outbreaks in the 21st century. *PloS Pathog.* 11, e1004804. doi: 10.1371/journal.ppat.1004804

Magill, S. S., O'leary, E., Janelle, S. J., Thompson, D. L., Dumyati, G., Nadle, J., et al. (2018). Changes in prevalence of health care-associated infections in U.S. hospitals. *N Engl. J. Med.* 379, 1732–1744. doi: 10.1056/NEJMoa1801550

Magruder, M., Sholi, A. N., Gong, C., Zhang, L. S., Edusei, E., Huang, J., et al. (2019). Gut uropathogen abundance is a risk factor for development of bacteriuria and urinary tract infection. *Nat. Commun.* 10, 5521. doi: 10.1038/s41467-019-13467-w

Mathy, V., Chousterman, B., Munier, A. L., Cambau, E., Jacquier, H., and De Ponfilly, G. P. (2020). First reported case of postneurosurgical ventriculoperitonitis due to kocuria rhizophila following a ventriculoperitoneal shunt placement. *Infect. Dis. Clin. Pract.* 28, 169–170. doi: 10.1097/IPC.0000000000000829

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

Mikosz, C. A., Smith, R. M., Kim, M., Tyson, C., Lee, E. H., Adams, E., et al. (2014). Fungal endophthalmitis associated with compounded products. *Emerg. Infect. Dis.* 20, 248–256. doi: 10.3201/eid2002.131257

Mody, L., Greene, M. T., Meddings, J., Krein, S. L., Mcnamara, S. E., Trautner, B. W., et al. (2017). A national implementation project to prevent catheter-associated urinary tract infection in nursing home residents. *JAMA Intern. Med.* 177, 1154–1162. doi: 10.1001/jamainternmed.2017.1689

Nouraei, H., Pakshir, K., Zareshahrabadi, Z., and Zomorodian, K. (2020). High detection of virulence factors by candida species isolated from bloodstream of patients with candidemia. *Microb. Pathog.* 149, 104574. doi: 10.1016/j.micpath.2020.104574

Odabasi, Z., and Mert, A. (2020). Candida urinary tract infections in adults. *World J. Urol* 38, 2699–2707. doi: 10.1007/s00345-019-02991-5

Oliveira, M., and Azevedo, L. (2022). Molecular markers: An overview of data published for fungi over the last ten years. *J. Fungi (Basel)* 8, 803. doi: 10.3390/jof8080803

Pappas, P. G., Lionakis, M. S., Arendrup, M. C., Ostrosky-Zeichner, L., and Kullberg, B. J. (2018). Invasive candidiasis. *Nat. Rev. Dis. Primers* 4, 18026. doi: 10.1038/nrdp.2018.26

Park, J. H., Oh, J., Sang, H., Shrestha, B., Lee, H., Koo, J., et al. (2019). Identification and antifungal susceptibility profiles of cyberlindnera fabianii in Korea. *Mycobiology* 47, 449–456. doi: 10.1080/12298093.2019.1651592

Pearson-Stuttard, J., Blundell, S., Harris, T., Cook, D. G., and Critchley, J. (2016). Diabetes and infection: assessing the association with glycaemic control in population-based studies. *Lancet Diabetes Endocrinol.* 4, 148–158. doi: 10.1016/S2213-8587(15)00379-4

Pfaller, M. A., Diekema, D. J., Turnidge, J. D., Castanheira, M., and Jones, R. N. (2019). Twenty years of the SENTRY antifungal surveillance program: Results for *Candida* species from 1997-2016. *Open Forum Infect. Dis.* 6, S79–S94. doi: 10.1093/ofid/ofy358

Pulcrano, G., Roscetto, E., Iula, V. D., Panellis, D., Rossano, F., and Catania, M. R. (2012). MALDI-TOF mass spectrometry and microsatellite markers to evaluate candida parapsilosis transmission in neonatal intensive care units. *Eur. J. Clin. Microbiol. Infect. Dis.* 31, 2919–2928. doi: 10.1007/s10096-012-1642-6

Reiss, E., Tanaka, K., Bruker, G., Chazalet, V., Coleman, D., Debeaupuis, J. P., et al. (1998). Molecular diagnosis and epidemiology of fungal infections. *Med. Mycol* 36 Suppl 1, 249–257.

Rhodes, J., Abdolrasouli, A., Farrer, R. A., Cuomo, C. A., Aanensen, D. M., Armstrong-James, D., et al. (2018). Genomic epidemiology of the UK outbreak of the emerging human fungal pathogen candida auris. *Emerg. Microbes Infect.* 7, 43. doi: 10.1038/s41426-018-0098-x

Stielow, J. B., Levesque, C. A., Seifert, K. A., Meyer, W., Iriny, L., Smits, D., et al. (2015). One fungus, which genes? development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia* 35, 242–263. doi: 10.3767/003158515X689135

Svobodova, L., Bednarova, D., Ruzicka, F., Chrenkova, V., Dobias, R., Mallatova, N., et al. (2016). High frequency of candida fabianii among clinical isolates biochemically identified as candida pelliculosa and candida utilis. *Mycoses* 59, 241–246. doi: 10.1111/myc.12454

Teke, L., Baris, A., and Bayraktar, B. (2021). Comparative evaluation of the bruker biotyper and vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) systems for non-albicans candida and uncommon yeast isolates. *J. Microbiol. Methods* 185, 106232. doi: 10.1016/j.mimet.2021.106232

Thomaz, D. Y., Del Negro, G. M. B., Ribeiro, L. B., Da Silva, M., Carvalho, G., Camargo, C. H., et al. (2022). A Brazilian inter-hospital candidemia outbreak caused by fluconazole-resistant candida parapsilosis in the COVID-19 era. *J. Fungi (Basel)* 8, 100. doi: 10.3390/jof8020100

Treguier, P., David, M., Gargala, G., Camus, V., Stamatoullas, A., Menard, A. L., et al. (2018). Cyberlindnera jadinii (teleomorph candida utilis) candidaemia in a patient with aplastic anaemia: a case report. *JMM Case Rep.* 5, e005160. doi: 10.1099/jmmcr.0.005160

Van Rijswijck, I. M. H., Van Mastrigt, O., Pijffers, G., Wolkers-Rooijackers, J. C. M., Abee, T., Zwietering, M. H., et al. (2019). Dynamic modelling of brewers' yeast and cyberlindnera fabianii co-culture behaviour for steering fermentation performance. *Food Microbiol.* 83, 113–121. doi: 10.1016/j.fm.2019.04.010

Wallace, R. J.Jr., Brown, B. A., and Griffith, D. E. (1998). Nosocomial outbreaks/pseudo-outbreaks caused by nontuberculous mycobacteria. *Annu. Rev. Microbiol.* 52, 453–490. doi: 10.1146/annurev.micro.52.1.453

World Health Organization (2022). *WHO fungal priority pathogens list to guide research, development and public health action.* WHO. Available at: https://www.who.int/publications/i/item/9789240060241.

Xiao, M., Chen, S. C., Kong, F., Xu, X. L., Yan, L., Kong, H. S., et al. (2020). Distribution and antifungal susceptibility of candida species causing candidemia in China: An update from the CHIF-NET study. *J. Infect. Dis.* 221, S139–S147. doi: 10.1093/infdis/jiz573

Xiao, M., Wang, H., Lu, J., Chen, S. C., Kong, F., Ma, X. J., et al. (2014). Three clustered cases of candidemia caused by candida quercitrusa and mycological characteristics of this novel species. *J. Clin. Microbiol.* 52, 3044–3048. doi: 10.1128/JCM.00246-14

Yadav, A., Singh, A., Wang, Y., Haren, M. H. V., Singh, A., De Groot, T., et al. (2021). Colonisation and transmission dynamics of candida auris among chronic respiratory diseases patients hospitalised in a chest hospital, Delhi, India: A comparative analysis of whole genome sequencing and microsatellite typing. *J. Fungi (Basel)* 7, 81. doi: 10.3390/jof7020081

Zhai, B., Ola, M., Rolling, T., Tosini, N. L., Joshowitz, S., Littmann, E. R., et al. (2020). High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis. *Nat. Med.* 26, 59. doi: 10.1038/s41591-019-0709-7

Zhang, L., Xiao, M., Arastehfar, A., Ilkit, M., Zou, J., Deng, Y., et al. (2021). Investigation of the emerging nosocomial wickerhamomyces anomalus infections at a Chinese tertiary teaching hospital and a systemic review: Clinical manifestations, risk factors, treatment, outcomes, and anti-fungal susceptibility. *Front. Microbiol.* 12, 744502. doi: 10.3389/fmicb.2021.744502

Zhang, L., Yu, S. Y., Chen, S. C., Xiao, M., Kong, F., Wang, H., et al. (2020). Molecular characterization of candida parapsilosis by microsatellite typing and emergence of clonal antifungal drug resistant strains in a multicenter surveillance in China. *Front. Microbiol.* 11. doi: 10.3389/fmicb.2020.01320

TYPE Perspective
PUBLISHED 20 April 2023
DOI 10.3389/fpubh.2023.1177695

Check for updates

OPEN ACCESS

EDITED BY
Peera Hemarajata,
Los Angeles County Department of Public Health, United States

REVIEWED BY
Emily Smith,
Theiagen Genomics,
United States
John Bell,
California Department of Public Health,
United States

*CORRESPONDENCE
Paige M.K. Larkin
✉ paigemklarkin@gmail.com

SPECIALTY SECTION
This article was submitted to Infectious Diseases: Epidemiology and Prevention, a section of the journal Frontiers in Public Health

RECEIVED 01 March 2023
ACCEPTED 30 March 2023
PUBLISHED 20 April 2023

CITATION
Tartanian AC, Mulroney N, Poselenzny K, Akroush M, Unger T, Helseth DL Jr, Sabatini LM, Bouma M and Larkin PMK (2023) NGS implementation for monitoring SARS-CoV-2 variants in Chicagoland: An institutional perspective, successes and challenges.
Front. Public Health 11:1177695.
doi: 10.3389/fpubh.2023.1177695

NGS implementation for monitoring SARS-CoV-2 variants in Chicagoland: An institutional perspective, successes and challenges

Aileen C. Tartanian[1], Nicole Mulroney[1], Kelly Poselenzny[1], Michael Akroush[1], Trevor Unger[1], Donald L. Helseth Jr.[1], Linda M. Sabatini[1,2], Michael Bouma[1] and Paige M.K. Larkin[1,2]*

[1]NorthShore University HealthSystem, Evanston, IL, United States, [2]Pritzker School of Medicine, The University of Chicago, Chicago, IL, United States

Identification of SARS-CoV-2 lineages has shown to provide invaluable information regarding treatment efficacy, viral transmissibility, disease severity, and immune evasion. These benefits provide institutions with an expectation of high informational upside with little insight in regards to practicality with implementation and execution of such high complexity testing in the midst of a pandemic. This article details our institution's experience implementing and using Next Generation Sequencing (NGS) to monitor SARS-CoV-2 lineages in the northern Chicagoland area throughout the pandemic. To date, we have sequenced nearly 7,000 previously known SARS-CoV-2 positive samples from various patient populations (e.g., outpatient, inpatient, and outreach sites) to reduce bias in sampling. As a result, our hospital was guided while making crucial decisions about staffing, masking, and other infection control measures during the pandemic. While beneficial, establishing this NGS procedure was challenging, with countless considerations at every stage of assay development and validation. Reduced staffing prompted transition from a manual to automated high throughput workflow, requiring further validation, lab space, and instrumentation. Data management and IT security were additional considerations that delayed implementation and dictated our bioinformatic capabilities. Taken together, our experience highlights the obstacles and triumphs of SARS-CoV-2 sequencing.

KEYWORDS

SARS-CoV-2, sequencing, molecular microbiology, molecular diagnostics, next generating sequencing

Introduction

Next generation sequencing (NGS) has been pivotal for understanding the impact of SARS-CoV-2 variants on transmission, pathogenicity, disease severity, vaccine and therapy efficacy, and diagnostic detection (1). For instance, the Omicron variant has been shown to evade the immune response in patients previously infected with SARS-CoV-2 or vaccinated against SARS-CoV-2 (2–5), render the majority of monoclonal antibody therapies ineffective (4), and cause more infections in younger patients compared to other variants (6). Conversely, the Omicron variant was associated with a lower 28-day mortality, ICU admission rate, and oxygen requirements compared to Delta (7).

With the benefits of SARS-CoV-2 sequencing highlighted, implementing such testing is attractive to many healthcare systems, but there are numerous challenges and considerations that should be addressed. Here we describe our experience with implementing NGS for SARS-CoV-2 variant analysis at NorthShore University HealthSystem (NSUHS) molecular diagnostic laboratory (MDL). As a fully integrated healthcare system, NSUHS-Edward-Elmhurst Health (EEH) serves over 4.2 million residents across northeast Illinois, including the city of Chicago and six suburban counties. The system currently encompasses 8 hospitals and over 300 outpatient centers. The NSUHS MDL was the first clinical laboratory in Illinois to perform SARS-CoV-2 testing (8) and has performed over 800,000 SARS-CoV-2 diagnostic assays to date. Our initial goals of SARS-CoV-2 sequencing was to detect shifts and emergence of lineages in real time, but challenges with staffing, turn-around-time (TAT), and sample selection complicated this.

Demand on the lab and lab staff

The MDL has ample experience with NGS, given the breadth of oncology NGS assays performed, uniquely positioning the laboratory to bring in SARS-CoV-2 sequencing compared to laboratories without sequencing experience. We officially launched COVIDSeq on an Illumina NextSeq 550Dx (San Diego, CA, United States) in March of 2021 after delays due to installation, training, and reagent acquisition. Once launched, COVIDSeq productivity was constrained by the priority given to clinical diagnostic assays for staffing and freezer storage. Based on these factors, the percentage of SARS-CoV-2 samples tested by the MDL that progressed to sequencing ranged from 0 to 18.5% monthly for 2022, when sequencing was performed on a regular basis (Table 1). Samples were selected based on testing location and available media volume, with the exact number of samples tested fluctuating due to balancing the cost of a run and the availability of reagents and technologists to perform sequencing within a timeframe. Samples from 2020 and 2021 were run retrospectively, but due to delays discussed in detail below, most of 2021 was spent troubleshooting, validating, and optimizing. The use of a manual bioinformatics pipeline and analysis, as discussed below, complicated analysis. The addition of a new technologist to lead SARS-CoV-2 sequencing and move to more automation, both for the wet lab and dry lab components, facilitated more streamlined SARS-CoV-2 sequencing in 2022. Looking at 2022, there were factors that directly impacted the number of samples that could be sequenced per month. In September, we extracted and prepared libraries for 163 samples. However, our liquid handler malfunctioned by erroneously releasing all pipette tips and crashing the program. All samples had been depleted and the libraries were rendered unsavable. In November and December 2022, we exhausted our purchased sequencing reagents and did not have approval to order additional sequencing reagents due to the high cost that exceeded the allotted budget. All SARS-CoV-2 sequencing was self-funded by our institution, requiring careful planning and restricting the ability to expand sequencing capacity significantly.

Our health system utilizes several different RT-PCR platforms for SARS-CoV-2 testing, which supports large volume testing in a variety of settings, including point-of-care and at each of our hospitals. However, this also led to multiple different swabs, transport media, and sample volumes. These variations were due to different assay requirements, sporadic swab and transport media shortages, and testing locations stocking different swabs. Due to early implementation of SARS-CoV-2 RT-PCR testing, we performed testing for a number of outreach non-affiliated sites that used a variety of swabs. The utilization of multiple instruments, many without available cycle threshold (Ct) values prevented establishment and selection of samples with appropriate Ct values. Often, labs will set a minimum Ct value for sequencing to increase sequencing yield, but we did not have that ability given the lack of available Ct value data. With sequencing any positive sample in 2022, only 25.3–57.0% of positive samples resulted in a consensus sequence for a SARS-CoV-2 lineage (Table 1).

TABLE 1 Summary of SARS-CoV-2 samples tested clinically and sequenced at NSUHS for 2022.

Month (2022)	# SARS-CoV-2 clinically tested	# Positive SARS-CoV-2	# Positives sequenced	# Samples of sequenced with consensus sequence	% Samples of positives sequenced	% Samples of sequenced positives with a consensus sequence
January	25,892	6,062	607	272	10.0	44.8
February	20,396	1,523	281	71	18.5	25.3
March	24,870	1,133	134	35	11.8	26.1
April	30,301	3,065	258	147	8.4	57.0
May	33,106	5,963	432	211	7.2	48.8
June	24,988	4,423	433	170	9.8	39.3
July	23,230	4,556	482	252	10.6	52.3
August	22,467	3,674	299	119	8.1	39.8
September	22,467	2,461	0	N/A	0.0	N/A
October	24,972	2,394	297	75	12.4	25.3
November	28,592	2,908	0	N/A	0.0	N/A
December	27,763	3,857	0	N/A	0.0	N/A

Some explanations for this wide range include low viral load (most EUA platforms used for clinical testing at our healthcare system only provided positive/negative results without a Ct value), and user and instrument error, including issues described below when automation was implemented. Due to lack of staffing and available resources, we did not collect this data for 2021 as we had to retrospectively sequence samples.

Complications with manual processing

The lab performed SARS-CoV-2 library preparations and sequencing using Illumina's COVIDSeq™ RUO Test. As with most NGS assays, the library preparation portion of COVIDSeq is costly in terms of time and labor. Initially, all library preparations were performed manually, requiring two full 8 h shifts for one technologist to complete (Table 2). While there were two technologists trained on the COVIDSeq assay, these technologists also were performing molecular diagnostic assays for clinical care, limiting their ability to prepare libraries for COVIDSeq to once per week.

The increased demand for clinical plastic consumables caused backorders and supply chain issues, restricting our ability to regularly perform COVIDSeq. During the first year, it was difficult to acquire an adequate amount of pipette tips to perform not only COVIDSeq but any of our routine molecular diagnostic clinical assays. Manually performing one COVIDSeq library preparation would consume 33 pipette tip boxes (Table 2). Therefore, to sequence a full run of 384 samples, over 130 tip boxes would be required. Regular and rapid sequencing during shifts and emergence of lineages such as Alpha, Delta, and Omicron would have been beneficial as these results would contribute to the local and global sequencing effort as well as guide hospital policies (e.g., allowed meeting size). However, the clinical assays consumed the necessary pipette tips and other plastic consumables so this was not feasible.

Complications with automatic processing

With technologist time and consumables preciously scarce, two automated liquid handlers were purchased to supplement the labor demand required for this initiative. The PerkinElmer Janus G3 liquid handler (Waltham, MA, United States) was chosen to facilitate RNA extraction because it had a high volume capacity and was relatively easy to use. Unfortunately, several calibration corrections were required after initial install due to persistent issues with probe pressure and pipette tip compatibility resulting in inconsistent reagent and sample volumes. Most calibrations would require the onsite visit of a field service technician, delaying implementation even further. Once resolved, Janus was compatible with our already implemented ThermoFisher KingFisher Flex (Waltham, MA, United States) instrument for viral RNA extraction/purification.

The Beckman Coulter i7 liquid handler (Brea, CA, United States) was purchased to automate library preparation. Because COVIDSeq library preparations of 96 samples require two thermocyclers running in tandem, and the i7 had only one, the batches were halved from 96 to 48 to accommodate the missing thermocycler. The i7 reduced hands-on time from 2 days to 7 h. In addition, the i7 uses only 30% of the number of tip boxes (Table 2). While there was a greater supply of tips for the i7 compared to manual pipette tips, we went one step further to decrease our chances of competing with labs for tips by using the unusual pipette tip size of 190 μL.

The i7 is convenient and improves workflow, but there were many challenges in establishing this assay automation. For instance, hard shell 96-well plates were on backorder when the i7 arrived, so we used non-hardshell 96-well plates. These plates melted and warped from the heat of the thermocycler, causing the i7 to drop, crush, and toss the plates as the grippers attempted to move them. Similarly, these plates proved to be incompatible with the reusable lid used by the i7 thermocycler. During a run, the i7 would sense every time the plate was improperly sealed and stop the program. These problems required constant attention by our lab staff, manually adjusting the fit of the thermocycler lid. This persisted until the correct 96-well plates could be obtained.

Automated liquid handlers can be the source of numerous errors that are difficult to identify and troubleshoot. For example, we observed a consistent reduction in consensus sequence yield for samples positioned on the left side of the 96-well plates compared to the right side. After troubleshooting, the lab determined that the i7 instrument lacked steps within the run script to re-suspend magnetic beads prior to arraying them into samples. As the beads settled to the bottom of the source tube through the duration of the library preparation, the i7 would dispense bead storage buffer, absent of beads, to the samples on the left side of the plate while the samples on the right side received the majority of beads. Once the mixing step was supplemented into the run script, we noted an improved uniformity of sample performance coupled with vastly improved library concentration yield. Despite this fix, the library concentration from the i7 would remain inferior to the yield of manual library preparation. And, like the Janus, i7 calibrations, updates, and repairs would often be delayed because they required an onsite visit from a service technician.

We continued to identify opportunities for improved efficiency. The COVIDSeq program on our i7 calls for all reagents to be placed in 1.5 ml tubes and kept chilled on a cold block on the deck, reducing hands-on work for technologists, but it could take up to 3 h for a technologist to prepare the 1.5 ml tubes of master mixes. Over time, we reduced this timeframe by over 80% because we found that many of these master mixes could be prepared and frozen in advance without sacrificing library preparation performance (Table 2).

Manual library preparations would typically produce >150 nM pooled libraries, but switching to automation resulted in libraries of <8 nM, despite the corrections made to the run script. Pools with a molarity <0.5 nM would result in a total batch failure defined as a 0%

TABLE 2 Comparison of manual vs. automated library preparations.

	Manual (96 samples)	Automated (48 samples)
No. batches per sequencing run	4	8
No. of pipette tip boxes	33	5
Hands-on tech time	16 h	3 h→30 min
Turn around time	2 days	7 h
Final pool molarity	>150 nM	<8 nM

consensus sequence. However, above 0.5 nM, we found no correlation ($R^2 = 0.0442$) between a pool's molarity and the percentage of samples resulting in a consensus sequence (Supplementary Figure S1). Ideally, we would quantify all individual specimens after RNA extraction, cDNA synthesis, and library preparation, removing low-concentration samples at each QC step. However, our lab does not have a high throughput way to quantify 96 samples at a time, so we only quantified each batch's pooled library prior to combining the pools for sequencing.

Bioinformatics and cybersecurity

NGS generates millions of sequencing reads per sample, and analyzing these reads requires a robust bioinformatics pipeline in an effort to detect and track novel variants. When the bioinformatics infrastructure is insufficient to support this immense quantity of data, institutions typically opt for commercially available solutions; either cloud-based or local, for their bioinformatics pipeline needs due to ease of use and readily available customer support. Cloud-based applications have the benefit of ease-of-use and easily accessible vendor support; however, the ever-growing push for cloud application usage provides tremendous cybersecurity concern for institutions and often requires a lengthy and in-depth risk assessment, which can delay implementation. Using a local analysis bioinformatics application platform can reduce concern from a cybersecurity perspective, but it increases cost as these systems often require the purchase of licensed software and additional hardware.

Our initiative to implement a SARS-CoV-2 NGS assay was driven by immediate need to contribute in variant tracking within our community. Due to urgent importance and to avoid further delay in implementation, we opted to purchase the local Illumina DRAGEN server (as opposed to Illumina's cloud-based application BaseSpace) to be the primary source of our bioinformatics data analysis. At the time, a BaseSpace subscription would have forced an extensive, lengthy risk assessment by our cybersecurity team as these cloud-based applications do not always satisfy standard HIPAA requirements to protect personal health information (PHI).

The DRAGEN COVIDSeq test local pipeline provided a summary report of positive or negative results along with output directories containing the desired FASTA and VCF files. FASTAs, BAMs and VCFs generated by the Illumina DRAGEN software on the NextSeq 550Dx sequencer were copied to a separate Linux server for analysis. Initially, we ran our own variant calling pipeline using open source software (using samtools), visualizing the results in IGV, and running a local copy of Ensembl VEP for COVID-19 to annotate the variant consequences. This labor-intensive effort was quickly abandoned when we began using more specialized open source software packages provided by Nextstrain, Pangolin, and Nextclade, reducing the necessity of manual analysis. After using Nextstrain (9) for a few months, we recognized that variant nomenclature was evolving away from Nextstrain clade names to Pango lineage (10) and WHO labels. To generate Pango names we analyzed merged FASTA files using the latest version of Pangolin (11). Nextclade (12) was also used to compare and summarize variant classifications by uploading our merged FASTA files (13).

Launching the DRAGEN COVIDSeq local pipeline was initiated via the Linux command line terminal. This method is extremely foreign to users who are accustomed to GUI-based software with only little to moderate Linux command line experience. Customer support was a necessity, particularly support via vendor remote access, as we experienced frequent pipeline analysis failures along with connectivity issues between the DRAGEN server and the NextSeq 550Dx. Vendor bioinformatics support is generally equipped to support their customers remotely. NorthShore HIT did not permit vendor remote support access, limiting our only options to lengthy phone conversations or email correspondence. With restricted remote access to independently investigate and troubleshoot, vendors rely on these often mutually time consuming methods to investigate and eventually resolve the issue. Lack of proper vendor remote support to address these issues contributed to lengthy delays in data processing as resolution to these problems often extended across multiple days.

With this workflow, we quickly realized that our goal to track lineage shifts in real time would be extremely difficult to accomplish. Available bandwidth for our highly talented yet small bioinformatics team was limited, as our established clinical oncology NGS assays were beginning to rebound to pre-pandemic volumes. Building and maintaining a local pipeline intended to track current lineages shifts required a considerable amount of bioinformatics support beyond the limits of our available institutional resources.

Internal bioinformatics resources were not the only struggle experienced through this initial process. The laboratory workflow required for the DRAGEN COVIDSeq test pipeline included a requirement for a positive, negative, and no template control for each set of 96 indices to be included in each sequencing run. In the event of a control failure, the entire set of 96 samples became invalid. To avoid risk of control failure, each positive control required a fresh serial dilution prior to each library preparation. These dilutions were not recommended to be stored long term. Since our intent was to only sequence known SARS-CoV-2 positive samples, the inclusion of controls seemed to hold little value and only added complexity to the workflow.

The local DRAGEN COVIDSeq pipeline did provide some upside. Each analysis completed rather quickly (usually within 1 to 2 h) and provided the necessary output data required for lineage identification. However, because the workflow to maintain this pipeline became unmanageable, we made the decision to purchase a BaseSpace subscription and shift our analysis to this cloud-based application. This transition required a lengthy approval process through our HIT cybersecurity team as cloud-based NGS data analysis increases potential risk to loss of PHI. To diminish this risk, we decided that all samples would remain de-identified throughout the wet bench, sequencing, and post sequencing analysis. All data would be presented as aggregated de-identified data with no link to clinical information. We did not have permission from HIT to submit any data to GISAID. Not only were our samples de-identified on the sequencer, but our institution considers date collected as PHI. This information is requested by GISAID for submission. Clinical microbiology laboratories at other institutions were able to submit completely anonymized samples to their academic colleagues for sequencing and in turn were able to successfully report de-identified metadata to GISAID and NCBI (14). These labs had IRBs that allowed patient-level data to be reported back to public health entities as the clinical labs retained access to the patient-level data while the academic sequencing partners did not have access (14). This approach, which requires institutional approval, infrastructure for de-identifying and re-identifying, and access to academic sequencing laboratories, would be ideal to allow dissemination of data to public health and biorepositories. In our case, sequencing was so delayed that our public health colleagues would have already sequenced those samples of interest, creating another hurdle for rapid collaboration.

Using the DRAGEN COVIDSeq pipeline via BaseSpace Sequence Hub resolved many of the previously mentioned concerns, including ease of use. Although analysis times increased by four-fold due to the shared traffic of the cloud-based server, the data analysis process was exponentially easier as it required very little intervention from internal staff and remote support was easily available to resolve problems. However, when launching the DRAGEN COVID lineage application, the sample selection process seems to be the most taxing step. Samples can be selected in groups, but careful attention is required as it is easy to unintentionally include or exclude samples from analysis. Identifying samples to be analyzed through the application can be difficult as the sample list includes both completed and analysis-pending samples. These concerns are rather minor compared to our prior workflow and the DRAGEN COVID lineage application has provided a manageable data analysis workflow as the application provides mapping/alignment and variant calling features. Open source databases, like NextClade and Pangolin, are routinely updated and made available for analysis through the application, and the data is easily viewable and managed by multiple users.

Clinical relevance/discussion

Molecular diagnostic assays that directly impact patient care were prioritized over SARS-CoV-2 sequencing, posing a challenge to continue RUO sequencing at high capacity. This was particularly the case during SARS-CoV-2 waves, when staffing was reduced due to illness and supplies were in high demand (15). As a result, there were substantial delays (>1 month) in sequencing SARS-CoV-2 specimens, contrasting with our original plan of using COVIDSeq to capture shifts and emergence of SARS-CoV-2 lineages. In addition to the cost, sequencing all specimens would likely provide little additional information as most samples received during pandemic waves would have the same composition of lineages that would be better captured with a smaller representative sample selection. On the other hand, between waves, our sample volume was too low to form any statistically relevant conclusions. Furthermore, it would take substantial time to accumulate 384 specimens for a full sequencing run, delaying results or forcing a partial run, which was costly. Surges in cases led us to recruit additional resource staff and research lab team members to work additional shifts to propel sequencing efforts, manually sorting through the samples to confirm positives, creating specimen labels, aliquoting, and documenting.

As previously discussed, our results were de-identified and mass aggregated to demonstrates shifts and trends within our patient population. While ideally we could share our results with our local health department to aid in their sequencing efforts, our results were not only delayed, but also did not have linked clinical data. This meant that sequencing efforts were unnecessarily duplicated due to inability to coordinate and share results, furthering the documented gap between public health labs and clinical labs (1, 16). We were, however, able to capture data categorized by symptomatic vs. asymptomatic cases and had these samples designated with their own test code for easy sorting and comparison. This comparison relied on trusting that physicians selected the correct test code indicating the presence or absence of symptoms. While we had planned to use these data to make comparisons between lineage and symptomatic state, upon review, we found that a small portion of physicians erroneously ordered the wrong test code and thus, accurate conclusions required substantial review. If the test codes had been appropriately ordered, the comparison in lineage between symptomatic and asymptomatic patients could have contributed to our knowledge in the field. In the end, we were able to share monthly trends with our healthcare system, modeling what other institutions have done (14).

Our decisions to sequence various populations and ultimately switch to mostly inpatient and ED specimens likely resulted in selection bias toward patients whose SARS-CoV-2 infection was not only symptomatic, but severe enough to seek hospital treatment, as well as selecting toward patients from high risk ages (including infants and those over the age of 65 years old) and individuals with pre-existing health conditions. The challenges of inferring clinical impact of variants have been well documented (17) as it is impossible to get a truly representative sample. Severity of symptoms is subjective and testing restrictions fluctuated throughout the pandemic, with some hospital systems only allowing the sickest patients to get tested (17). Moreover, COVID-19 studies often focus on hospitalized patients, not representative of the general population (17). The shift to at-home antigen testing also biases against sequencing asymptomatic or mildly symptomatic patients (16).

Despite the issues discussed previously, our data were useful in a broader capacity for our healthcare system. While there were detected cases of Omicron in our state, our sequencing confirmed the presence of Omicron in our patient population. This contributed to discussions on policies for masking and permitted meeting sizes. Furthermore, in conjunction with *in silico* analysis, we used COVIDSeq to test detection of sequence-confirmed variants in our lab-developed SARS-CoV-2 assay. We were able to confirm that the primers for this clinical assay could still detect even the most recently detected lineages of SARS-CoV-2. This was a concern for clinical laboratories across the world as the Alpha and Omicron variants exhibited spike (S) gene dropouts on assays that detect the S gene (18). While we do not currently utilize any assay that targets S gene, mutations can occur in any region of the genome and thus it is important to monitor whether these mutations impact the ability for our assays to detect SARS-CoV-2.

The question of balancing cost, in terms of time and money, as well as staffing remains difficult and potentially unsustainable in the long-term for genomic surveillance. At times where multiple lineages are circulating, there was a push for more sequencing to better document lineage changes within our patient population, with the caveat that we do not have the capacity to provide rapid TAT for COVIDSeq. When there was an overwhelmingly predominant lineage, there was less institutional support for routine sequencing as, until a new variant of interest or concern is identified or mutations within current circulating variants would render treatments ineffective, the results would not impact hospital protocols. However, this approach would prevent detection of shifts in lineages as well as detection of novel lineages.

The challenges described here were not unique to our health care system. Both the importance of localized surveillance efforts as well as the extensive challenges in terms of labor force, supply chain issues, and coordinated data acquisition, analysis and sharing became painfully evident. In recognition, Congress passed the "Tracking Pathogen Act" as part of pandemic preparedness measures within the FY2023 Omnibus legislation. This Act directs the Department of Health and Human Services to issue guidance and to support such efforts.

Data availability statement

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

Author contributions

AT: writing—original draft, writing—review and editing, visualization, data curation, and investigation. NM, KP, MA, and TU: writing—review and editing, data curation, and investigation. DH: writing—review and editing, formal analysis, and software. LS: writing—review and editing, supervision, and conceptualization. MB: writing—conceptualization, original draft, writing—review and editing, supervision, data curation, and formal analysis. PL: writing—conceptualization, original draft, writing—review and editing, and project administration. All authors contributed to the article and approved the submitted version.

Acknowledgments

We extend our gratitude to the NSUHS MDL for all of their work providing timely and accurate patient results as well as assisting with COVIDSeq endeavors through coverage and testing support. We would also like to thank our administrative team, Karen Kaul, Lakshmi Halasyamani, and Matt Charles, for their support. Brian Staes and Mark Delamar were instrumental in positive specimen collection outreach and the NSUHS Core Laboratory lead the specimen receiving challenge. The NSUHS Microbiology Laboratory was essential in collection and storage of patient specimens. Donna Schora manually printed all labels and tracked positive specimens while the Research Laboratory stored all samples. We would also like to thank all resource individuals who contributed to the COVIDSeq project.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1177695/full#supplementary-material

SUPPLEMENTARY FIGURE S1

Lack of Correlation Between Library Pool Molarity and Percentage of Samples Generating a Consensus Sequence.

References

1. Martin MA, VanInsberghe D, Koelle K. Insights from SARS-CoV-2 sequences. *Science*. (2021) 371:466–7. doi: 10.1126/science.abf3995

2. Wang M, Zhou B, Fan Q, Zhou X, Liao X, Lin J, et al. Omicron variants escape the persistent SARS-CoV-2-specific antibody response in 2-year COVID-19 convalescents regardless of vaccination. *Emerg Microbes Infect*. (2023) 12:2151381. doi: 10.1080/22221751.2022.2151381

3. Dhama K, Nainu F, Frediansyah A, Yatoo MI, Mohapatra RK, Chakraborty S, et al. Global emerging omicron variant of SARS-CoV-2: impacts, challenges and strategies. *J Infect Public Health*. (2023) 16:4–14. doi: 10.1016/j.jiph.2022.11.024

4. Liu L, Iketani S, Guo Y, Chan JF, Wang M, Liu L, et al. Striking antibody evasion manifested by the omicron variant of SARS-CoV-2. *Nature*. (2022) 602:676–81. doi: 10.1038/s41586-021-04388-0

5. Ma C, Chen X, Mei F, Xiong Q, Liu Q, Dong L, et al. Drastic decline in sera neutralization against SARS-CoV-2 Omicron variant in Wuhan COVID-19 convalescents. *Emerg Microbes Infect*. (2022) 11:567–72. doi: 10.1080/22221751.2022.2031311

6. Lai A, Bergna A, Della Ventura C, Menzo S, Bruzzone B, Sagradi F, et al. Epidemiological and clinical features of SARS-CoV-2 variants circulating between April-December 2021 in Italy. *Viruses*. (2022) 14:2508. doi: 10.3390/v14112508

7. Beraud G, Bouetard L, Civljak R, Michon J, Tulek N, Lejeune S, et al. Impact of vaccination on the presence and severity of symptoms of hospitalised patients with an infection by the omicron variant (B.1.1.529) of the SARS-coV-2 (subvariant BA.1). *Clin Microbiol Infect*. (2022). doi: 10.1016/j.cmi.2022.12.020

8. Kaul K, Singh K, Sabatini L, Konchak C, McElvania E, Larkin P, et al. The value and institutional impact of an in-system laboratory testing during the COVID-19 pandemic. *Acad Pathol*. (2021) 8:23742895211010253. doi: 10.1177/23742895211010253

9. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407

10. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. (2020) 5:1403–7. doi: 10.1038/s41564-020-0770-5

11. O'Toole A, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol*. (2021) 7:veab064. doi: 10.1093/ve/veab064

12. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling, and quality control for viral genomes. *J Open Source Softw*. (2021) 6:3773. doi: 10.21105/joss.03773

13. Nextstrain. Clade assignment, mutation calling, and sequence quality checks. Available at: https://clades.nextstrain.org (Accessed 02/01/2023)

14. Wang J, Hawken SE, Jones CD, Hagan RS, Bushman F, Everett J, et al. Collaboration between clinical and academic laboratories for sequencing SARS-CoV-2 genomes. *J Clin Microbiol*. (2022) 60:e0128821. doi: 10.1128/jcm.01288-21

15. Mahilkar S, Agrawal S, Chaudhary S, Parikh S, Sonkar SC, Verma DK, et al. SARS-CoV-2 variants: impact on biological and clinical outcome. *Front Med*. (2022) 9:995960. doi: 10.3389/fmed.2022.995960

16. Ling-Hu T, Rios-Guzman E, Lorenzo-Redondo R, Ozer EA, Hultquist JF. Challenges and opportunities for global genomic surveillance strategies in the COVID-19 era. *Viruses*. (2022) 14:2532. doi: 10.3390/v14112532

17. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. (2020) 11:5749. doi: 10.1038/s41467-020-19478-2

18. McMillen T, Jani K, Robilotti EV, Kamboj M, Babady NE. The spike gene target failure (SGTF) genomic signature is highly accurate for the identification of alpha and omicron SARS-CoV-2 variants. *Sci Rep*. (2022) 12:18968. doi: 10.1038/s41598-022-21564-y

frontiers | Frontiers in Public Health

TYPE Correction
PUBLISHED 27 June 2023
DOI 10.3389/fpubh.2023.1234900

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Paige M. K. Larkin
✉ paigemklarkin@gmail.com

RECEIVED 05 June 2023
ACCEPTED 06 June 2023
PUBLISHED 27 June 2023

CITATION
Tartanian AC, Mulroney N, Poselenzny K, Akroush M, Unger T, Helseth DL Jr, Sabatini LM, Bouma M and Larkin PMK (2023) Corrigendum: NGS implementation for monitoring SARS-CoV-2 variants in Chicagoland: an institutional perspective, successes and challenges. *Front. Public Health* 11:1234900. doi: 10.3389/fpubh.2023.1234900

Corrigendum: NGS implementation for monitoring SARS-CoV-2 variants in Chicagoland: an institutional perspective, successes and challenges

Aileen C. Tartanian[1], Nicole Mulroney[1], Kelly Poselenzny[1], Michael Akroush[1], Trevor Unger[1], Donald L. Helseth Jr.[1], Linda M. Sabatini[1,2], Michael Bouma[1] and Paige M. K. Larkin[1,2]*

[1]NorthShore University HealthSystem, Evanston, IL, United States, [2]Pritzker School of Medicine, The University of Chicago, Chicago, IL, United States

KEYWORDS

SARS-CoV-2, sequencing, molecular microbiology, molecular diagnostics, next generating sequencing

A corrigendum on
NGS implementation for monitoring SARS-CoV-2 variants in Chicagoland: an institutional perspective, successes and challenges

by Tartanian, A. C., Mulroney, N., Poselenzny, K., Akroush, M., Unger, T., Helseth, D. L. Jr., Sabatini, L. M., Bouma, M., and Larkin, P. M. K. (2023) *Front. Public Health.* 11:1177695. doi: 10.3389/fpubh.2023.1177695

In the published article, an author name was incorrectly written as Aileen C. Tartaninan. The correct spelling is Aileen C. Tartanian.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

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

frontiers | Frontiers in Public Health

TYPE Case Report
PUBLISHED 09 May 2023
DOI 10.3389/fpubh.2023.1145733

Check for updates

OPEN ACCESS

EDITED BY
Varvara K. Kozyreva,
California Department of Public Health,
United States

REVIEWED BY
Xiao-Yong Zhan,
Sun Yat-sen University, China
Alina Ryabova,
Fritz Lipmann Institute (FLI), Germany

*CORRESPONDENCE
Hui Mao
✉ merrymh@yeah.net

†These authors have contributed equally to this work

RECEIVED 16 January 2023
ACCEPTED 11 April 2023
PUBLISHED 09 May 2023

CITATION
Du R, Feng Y, Wang Y, Huang J, Tao Y and Mao H (2023) Metagenomic next-generation sequencing confirms the diagnosis of *Legionella* pneumonia with rhabdomyolysis and acute kidney injury in a limited resource area: a case report and review.
Front. Public Health 11:1145733.
doi: 10.3389/fpubh.2023.1145733

Metagenomic next-generation sequencing confirms the diagnosis of *Legionella* pneumonia with rhabdomyolysis and acute kidney injury in a limited resource area: a case report and review

Rao Du[1†], Yinhe Feng[2†], Yubin Wang[1], Jifeng Huang[1], Yuhan Tao[1] and Hui Mao[1*]

[1]Department of Respiratory and Critical Care Medicine, West China Hospital, Sichuan University, Chengdu, China, [2]Department of Respiratory and Critical Care Medicine, Deyang People's Hospital, Affiliated Hospital of Chengdu College of Medicine, Deyang, China

Background: *Legionella* pneumonia, rhabdomyolysis, and acute kidney injury are called the *Legionella* triad, which is rare and associated with a poor outcome and even death. Early diagnosis and timely treatment are essential for these patients.

Case presentation: A 63-year-old man with cough, fever, and fatigue was initially misdiagnosed with common bacterial infection and given beta-lactam monotherapy but failed to respond to it. Conventional methods, including the first *Legionella* antibody test, sputum smear, and culture of sputum, blood, and bronchoalveolar lavage fluid (BALF) were negative. He was ultimately diagnosed with a severe infection of *Legionella pneumophila* by metagenomics next-generation sequencing (mNGS). This patient, who had multisystem involvement and manifested with the rare triad of *Legionella* pneumonia, rhabdomyolysis, and acute kidney injury, finally improved after combined treatment with moxifloxacin, continuous renal replacement therapy, and liver protection therapy.

Conclusion: Our results showed the necessity of early diagnosis of pathogens in severe patients, especially in Legionnaires' disease, who manifested with the triad of *Legionella* pneumonia, rhabdomyolysis, and acute kidney injury. mNGS may be a useful tool for Legionnaires' disease in limited resource areas where urine antigen tests are not available.

KEYWORDS

***Legionella* pneumonia, rhabdomyolysis, acute kidney injury, metagenomic next-generation sequencing, case report**

Introduction

Legionella pneumophila (*L. pneumophila*), a species of the *Legionella* genus, is the causative agent of Legionellosis, which contains two forms: the non-pneumonic form (Pontiac fever) and the acute pneumonic form (Legionnaires' disease) (1). Pontiac fever is an influenza-like syndrome. Legionnaires' disease is more severe and can be involved in extrapulmonary manifestation with severe pneumonia, which requires hospitalization and most commonly intensive care. The mortality was 4.6% in medical wards compared with 23.1% in the intensive care unit (ICU) (2). Extrapulmonary manifestations included abdominal pain, diarrhea of the digestive system, weakness, and fatigue of the musculoskeletal system, myoglobinuria of the urinary system, malaise of the nervous system, etc. The triad of *Legionella* pneumonia, rhabdomyolysis, and acute kidney injury (AKI) was rare, and its mortality was much higher than in patients who manifested only with *Legionella* pneumonia. Early diagnosis and prompt treatment is critical for such patients. However, the existing diagnosis method of legionellosis cannot meet the clinical demands.

Metagenomic next-generation sequencing (mNGS), a culture-independent method, can detect all pathogens from one specimen (3) and has been recommended by expert consensus for diagnosing challenging cases of complicated infectious disease (4). It is especially suitable for suspected infectious diseases with negative conventional methods. Herein, we present a case of *Legionella pneumophilia* infection that manifested initially as cough, fever, and fatigue, and led to a rare triad of *Legionella* pneumonia, rhabdomyolysis, and AKI, which was ultimately diagnosed by mNGS.

Case presentation

A 63-year-old man presented to the emergency department with a 3-day history of cough, fever, and fatigue. He was receiving piperacillin-tazobactam monotherapy at a local hospital. However, the patient failed to respond to the treatment. Initial vital signs were body temperature (>40°C), pulse 148 beats per minute, and blood pressure 146/73 mmHg. Although he had been administered 10 L of oxygen via a nasal cannula, the peripheral oxygen saturation was only 87%. Physical examination revealed somnolent consciousness, and auscultation revealed decreased breath sounds and scattered rales in both lower lobes. The muscle strength of the upper limbs was grade four, and that of the lower limbs was grade two. The initial laboratory test results are presented in Table 1. Urine analysis showed hematuria. The first serum antibody test of anti-*Legionella* was negative. DNA tests for chlamydia and mycoplasma were negative, as well as viral pharyngeal swabs for influenza A and B. Electrocardiography revealed sinus tachycardia. Computed tomography (CT) revealed extensive consolidation in both the lower lobes (Figures 1A, B). Invasive intubation and continuous renal replacement therapy (CRRT) were initiated immediately after admission, and intravenous meropenem (1,000 mg q8h) was given on the first day. Given that the patient had confusion, hyponatremia, elevated creatine kinase (CK), and severe pneumonia, Legionnaire's disease was suspected. Thus, moxifloxacin (400 mg, once a day) was added the next day for treatment. He was admitted to the intensive care unit (ICU) for the management of acute respiratory failure, massive rhabdomyolysis, and AKI.

TABLE 1 Laboratory analysis at admission.

Laboratory analysis	Level	Normal range
WBC	11.65 × 10^9/L	3–10 × 10^9/L
NEU%	94.0%	40–75%
PCT	>100 ng/ml	<0.046 ng/ml
CRP	295.00 mg/L	<5 mg/L
IL-6	111.00 pg/ml	0–7.00 pg/ml
ALT	173 IU/L	<40 IU/L
AST	571 IU/L	<35 IU/L
LDH	1,213 IU/L	120–250 IU/L
CK	27,848 IU/L	20–140 IU/L
Myoglobin	>3000.00 ng/mL	<58.0 ng/mL
CK-MB	25.12 ng/mL	<2.88 ng/mL
TPN-T	73.0 ng/L	0–14 ng/L
serum creatinine	616.00 μmol/L	48–79 μmol/L
BUN	16.5 mmol/L	2.6–7.5 mmol/L
eGFR	19.70 ml/min/1.73m^2	>90 ml/min/1.73m^2
Na^+	130.9 mmol/L	137.0–147.0 mmol/L
Glucose	24.0 mmol/L	3.90–5.90 mmol/L
Glycosylated hemoglobin	13.3%	<6.0%

WBC, white blood cell; NEU, neutrophil; ALT, alanine transaminase; AST, aspartate transaminase; LDH, lactate dehydrogenase; CK, creatine kinase; TPN-T, troponin T; BUN, blood urea nitrogen; eGFR, estimated glomerular filtration rate; Na, sodium.

Diagnostic assessment

The patient's sputum smear and sputum, blood, and bronchoalveolar lavage fluid (BALF) conventional common bacteria cultures failed to reveal a pathogen. Therefore, BALF and lung tissue samples were sent for mNGS analysis (BGI-500, Chengdu, China). The BGI company conducted mNGS as described previously (5), using samples of 0.5–3 mL sputum or lung tissue collected following standard procedures. The sample was agitated at 2,800–3,200 rpm for 30 min in a 1.5 mL microcentrifuge tube. DNA was extracted from a new tube of 0.3 mL sample tube using the TIANamp Micro DNA Kit (DP316; Tiangen Biotech, http://tiangen.com) according to the manufacturer's recommendations. DNA libraries were constructed using DNA fragmentation, end repair, adapter ligation, and PCR amplification. It generated high-quality sequencing data from sequencing libraries using the BGISEQ-500 platform. Subsequently, computational subtraction of the human host sequences (hg19) was performed and low-quality and short reads (<50 bp) were removed by Burrows-Wheeler Alignment, as well as low-complexity reads, the remaining data were classified by aligning them to BGI self-established database downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/), which contains 1,798 whole genome sequences of DNA viral taxa, 6,350 bacterial genomes or scaffolds, 1,064 fungi related to human infection, and 234 parasites associated with human diseases.

Abbreviations: *L. pneumophila, Legionella pneumophila*; AKI, acute kidney injury; BALF, bronchoalveolar lavage fluid; CK, creatine kinase; BCYE, buffered charcoal yeast extract; LEV, levofloxacin; mNGS, metagenomic next-generation sequencing; MOX, moxifloxacin.

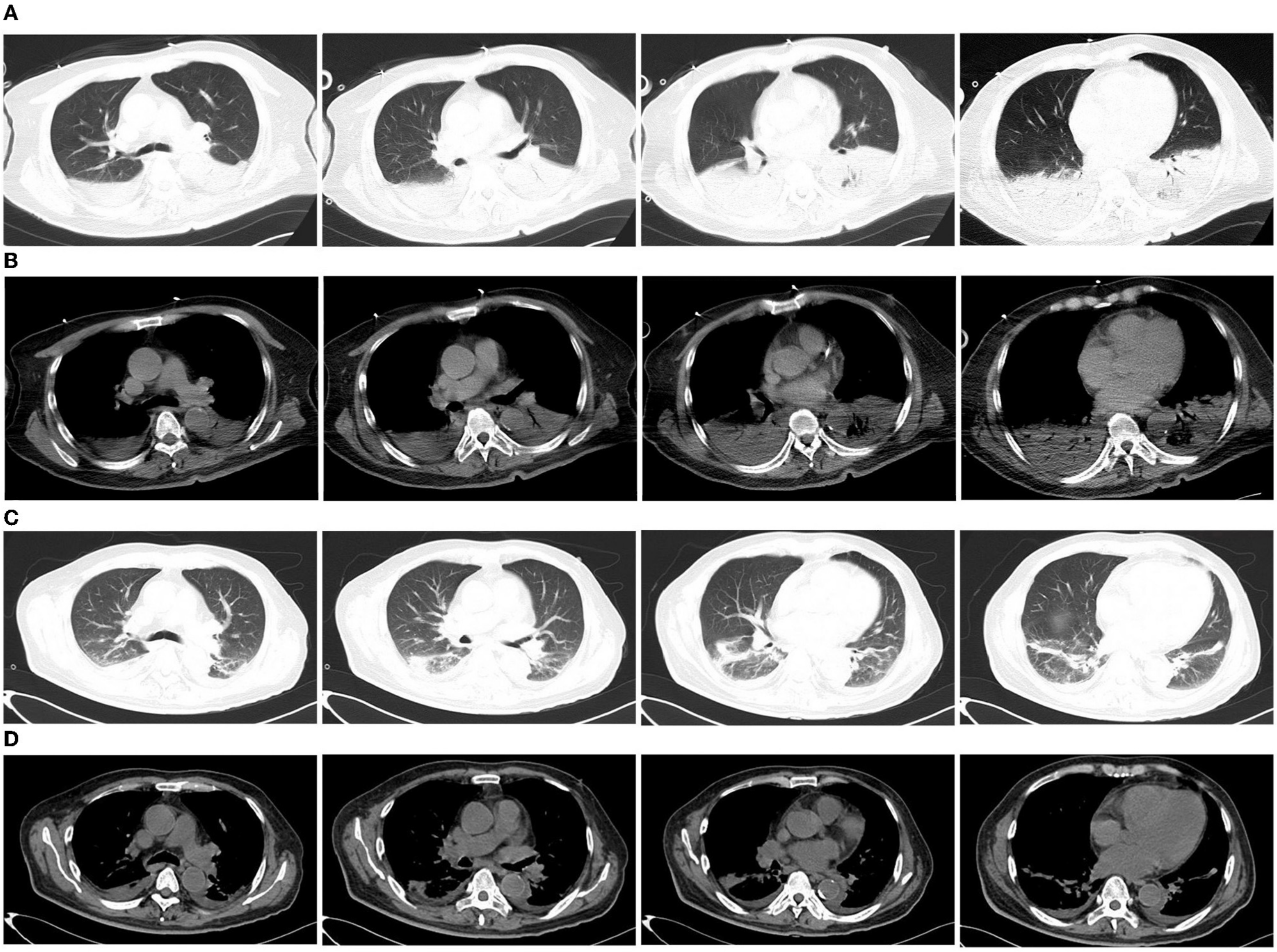


FIGURE 1
Chest computed tomography (CT) of the patient. **(A, B)** Chest CT on the first day of admission to the emergency department showed consolidation of the lower lobes of both lungs; **(C, D)** Chest CT 23 days after moxifloxacin treatment showed that the lesions were significantly absorbed.

In total, 147,673,559 clean sequence reads were obtained in the BALF. When human host reads were excluded, 1,021 sequence reads were identified as *Legionella* at the genus level, 980 of which matched *L. pneumophila* at the species level. mNGS of the lung tissue yielded 131,264,566 clean sequence reads. When human host reads were excluded, 309 sequence reads were identified as *Legionella* at the genus level, 291 of which matched *L. pneumophila* at the species level (Figure 2). Experts in respiratory illness, microbiology, and radiology interpreted the results to identify potential etiological agents. Moxifloxacin monotherapy was used according to these results, and meropenem was weaned. Chest radiography showed significant improvement after 7 days, and the patient was extubated. After 15 days in the ICU, the patient was transferred to the general ward for treatment. Kidney function recovered gradually, and the frequency of hemodialysis changed from once a day to every other day and continued until day 15. The second anti-*Legionella* antibody test was positive 22 days after the onset of the disease. After 23 days of moxifloxacin treatment, chest CT revealed significant absorption of lesions (Figures 1C, D). He was discharged from the hospital after 38 days with normal creatinine and CK, ALT, AST, LDH, WBC count, PCT, and CRP levels (Figure 3). The timeline of patients with relevant data on episodes and interventions is presented in Figure 4.

Discussion

Legionella is gram-negative rod-shaped bacteria ubiquitously found in fresh water environments and moist soil (6). *Legionella* pneumonia accounts for 2–15% of all community-acquired pneumonia (CAP) cases requiring hospitalization. It is the second most common cause of serious pneumonia and requires ICU admission (7). Legionnaire's disease is a severe form of pneumonia caused by *Legionella*; it can have extrapulmonary manifestations (8). Most cases are caused by *L. pneumophila*, whereas some are caused by *Legionella longbeachae*. Male sex (>50 years of age), smoking, and diabetes mellitus are the risk factors (9). Approximately 62% of the cases occur during summer and early autumn (10), which is related to the use of air conditioning. In this case, although the patient had no history of diabetes, his random blood glucose level was 24.0 mmol/L, and his glycosylated hemoglobin level was 13.3%. Therefore, the patient was diagnosed

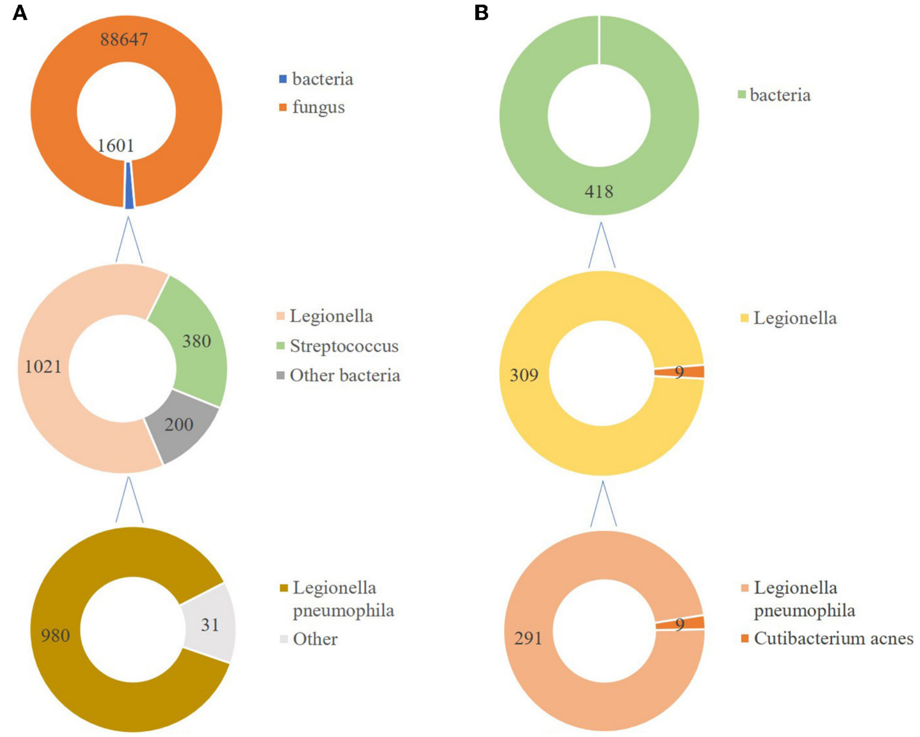
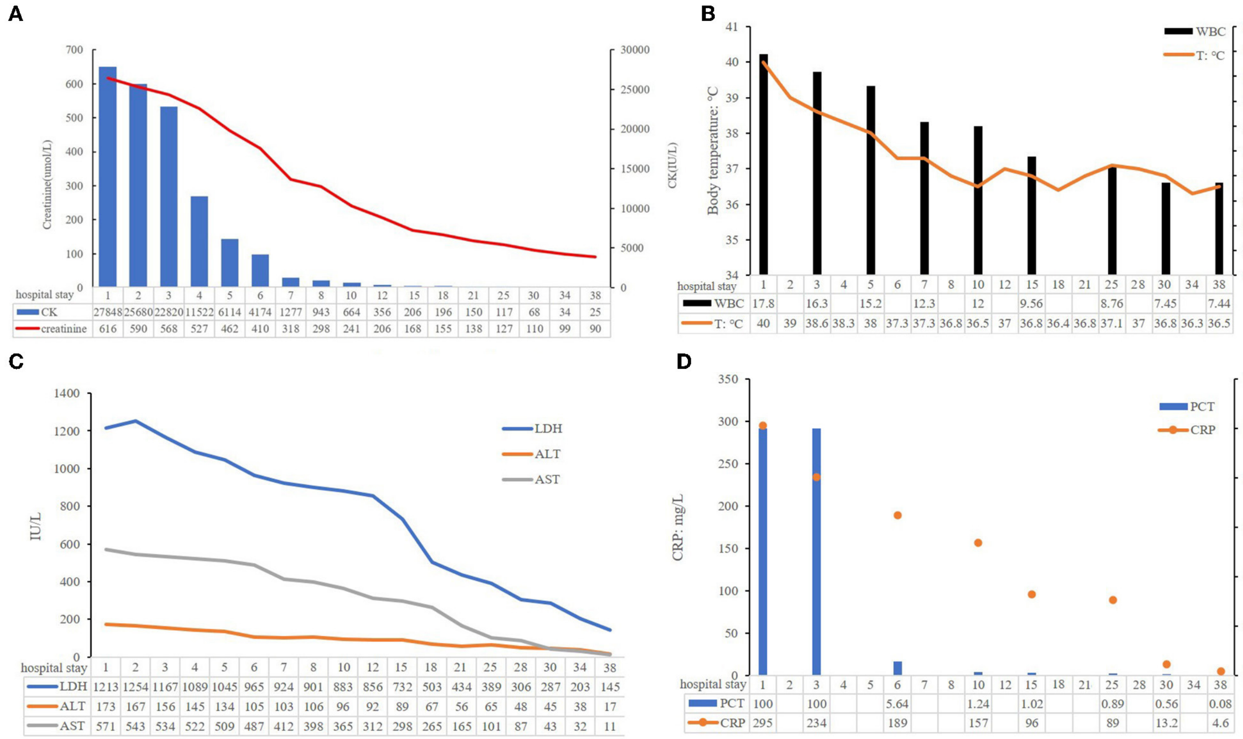


FIGURE 2
The corresponding reads of detected microorganism in **(A)** BALF and **(B)** lung tissue.



hospital stay	1	2	3	4	5	6	7	8	10	12	15	18	21	25	30	34	38
CK	27848	25680	22820	11522	6114	4174	1277	943	664	356	206	196	150	117	68	34	25
creatinine	616	590	568	527	462	410	318	298	241	206	168	155	138	127	110	99	90

hospital stay	1	2	3	4	5	6	7	8	10	12	15	18	21	25	28	30	34	38
WBC	17.8		16.3		15.2		12.3		12		9.56			8.76		7.45		7.44
T: °C	40	39	38.6	38.3	38	37.3	37.3	36.8	36.5	37	36.8	36.4	36.8	37.1	37	36.8	36.3	36.5

hospital stay	1	2	3	4	5	6	7	8	10	12	15	18	21	25	28	30	34	38
LDH	1213	1254	1167	1089	1045	965	924	901	883	856	732	503	434	389	306	287	203	145
ALT	173	167	156	145	134	105	103	106	96	92	89	67	56	65	48	45	38	17
AST	571	543	534	522	509	487	412	398	365	312	298	265	165	101	87	43	32	11

hospital stay	1	2	3	4	5	6	7	8	10	12	15	18	21	25	28	30	34	38
PCT	100		100			5.64			1.24		1.02			0.89		0.56		0.08
CRP	295		234			189			157		96			89		13.2		4.6

FIGURE 3
Some clinical indicators of this patient during hospitalization. **(A)** Creatinine and CK levels; **(B)** ALT, AST, and LDH levels; **(C)** WBC and body temperature; **(D)** PCT, CRP levels.

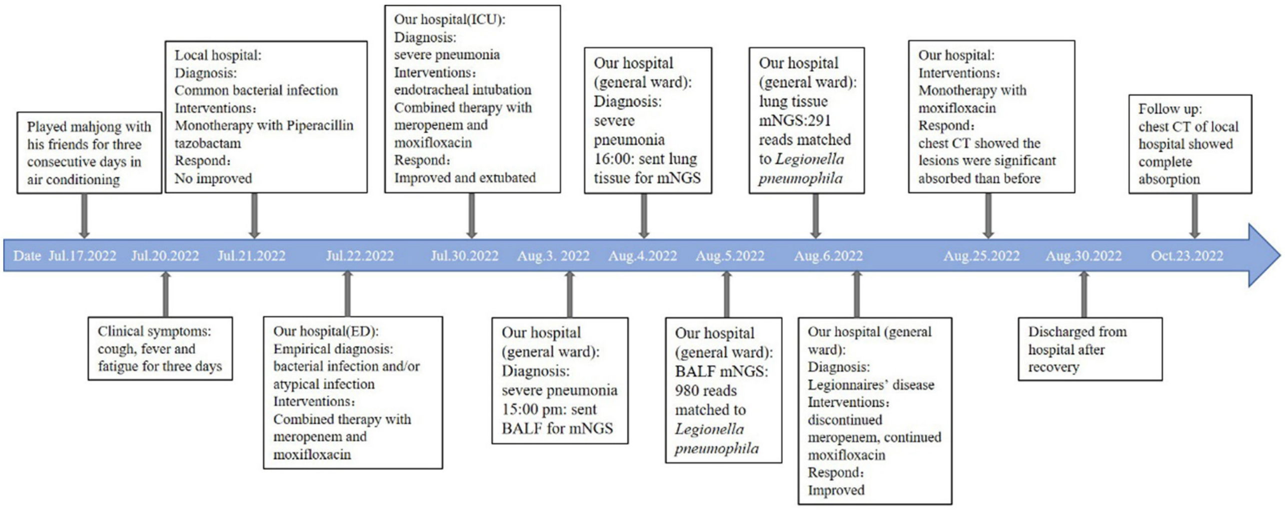

Played mahjong with his friends for three consecutive days in air conditioning
Local hospital: Diagnosis: Common bacterial infection Interventions: Monotherapy with Piperacillin tazobactam Respond: No improved
Our hospital(ICU): Diagnosis: severe pneumonia Interventions: endotracheal intubation Combined therapy with meropenem and moxifloxacin Respond: Improved and extubated
Our hospital (general ward): Diagnosis: severe pneumonia 16:00: sent lung tissue for mNGS
Our hospital (general ward): lung tissue mNGS:291 reads matched to *Legionella pneumophila*
Our hospital: Interventions: Monotherapy with moxifloxacin Respond: chest CT showed the lesions were significant absorbed than before
Follow up: chest CT of local hospital showed complete absorption
Date Jul.17.2022
Jul.20.2022
Jul.21.2022
Jul.22.2022
Jul.30.2022
Aug.3. 2022
Aug.4.2022
Aug.5.2022
Aug.6.2022
Aug.25.2022
Aug.30.2022
Oct.23.2022
Clinical symptoms: cough, fever and fatigue for three days
Our hospital(ED): Empirical diagnosis: bacterial infection and/or atypical infection Interventions: Combined therapy with meropenem and moxifloxacin
Our hospital (general ward): Diagnosis: severe pneumonia 15:00 pm: sent BALF for mNGS
Our hospital (general ward): BALF mNGS: 980 reads matched to *Legionella pneumophila*
Our hospital (general ward): Diagnosis: Legionnaires' disease Interventions: discontinued meropenem, continued moxifloxacin Respond: Improved
Discharged from hospital after recovery

FIGURE 4
Timeline of the patient with relevant data of the episodes and interventions.

with type 2 diabetes. Generally, he had four risk factors: sex, age, smoking history, and type 2 diabetes. Before the symptoms appeared, he played Mahjong with his friends for 3 consecutive days on air conditioning. We suspected that Mahjong parlor was the source of the *L. pneumophila* infection.

No clinical manifestations unique to Legionnaire's disease were observed. The symptoms of Legionnaires' disease from most to least common include fever, cough, chills, dyspnea, neurological abnormalities, myalgia or arthralgia, diarrhea, chest pain, headache, and nausea or vomiting (10). Non-specific laboratory findings are common, such as elevated CK level, myoglobinuria, hyponatremia, microscopic hematuria, and leukocytosis (10). The rate of hyponatremia was 25.3% in a study of CAP (11). In Legionnaires' disease, the rate of hyponatremia was 44.4% (12). Hyponatremia has a negative impact on multiple outcomes, such as the need for mechanical ventilation and ICU care, the duration of hospital or ICU stay. In particular, hyponatremia adds more than 10,000 RMB to the cost of care (13). In studies of *Legionella*-related CAP, hyponatremia (<133 mmol/L) was one of the strongest predictors (14, 15). Our patient had hyponatremia, elevated CK level, myoglobinuria, and leukocytosis, which were important reasons that the attending physician suspected *Legionella*-related CAP and administered moxifloxacin.

Furthermore, rhabdomyolysis is a syndrome caused by the breakdown and necrosis of muscle tissue and release of intracellular contents into the bloodstream (16). A diagnosis can be made when the serum CK level is >1,000 U/L (16). In adults, common causes are trauma and infection. The reported viruses and bacteria that can cause rhabdomyolysis include the following: influenza A and B, coxsackievirus, Epstein–Barr virus, primary human immunodeficiency virus, *Legionella* species, *Streptococcus pyogenes*, *Staphylococcus aureus* (*pyomyositis*), and *Clostridium* (17). Rhabdomyolysis caused by bacteria is associated with high mortality and morbidity: 57% of the cases lead to AKI, and 38% result in death (18). It can cause subacute- or acute-onset myalgia, transient muscle weakness, and dark tea- or cola-colored urine (16). Rhabdomyolysis induced by *Legionella* is rare. Prompt recognition is important for doctors to provide timely and appropriate treatment. In the present case, the patient was initially administered piperacillin-tazobactam monotherapy; however, no response was observed. Differential diagnoses of *S. pyogenes* and *S. aureus* can be excluded.

AKI refers to a sudden loss of excretory kidney function determined by increased serum creatinine levels and reduced urinary output, which can be caused by various factors (19). Infections and hypovolemic shock are the primary causes of AKI in low- and middle-income countries (19). In high-income countries, it mostly occurs in hospitalized older patients and is associated with sepsis, drug use, or invasive procedures (19). In non-traumatic rhabdomyolysis, AKI related to myoglobinuria is a serious complication. Patients who develop AKI have an increased mortality rate of 80% (20). Considering *Legionella* infection, the exact pathophysiology of rhabdomyolysis and AKI is poorly understood and is currently suspected to be endotoxin-mediated (21). It is thought to be induced by rhabdomyolysis; it is also thought to be induced by direct bacterial inoculation of the renal tissue (22). In our case, the CK level of the patient was 27,848 IU/L, he was anuric, with no more than 40 mL dark-tea urine in 24 h, and the synchronous serum creatinine was 616.00 μmol/L when he came to our hospital. In this patient, renal function was severely impaired in the initial stage of the disease, and it was difficult to determine whether AKI was indirectly associated with *L. pneumophila* via rhabdomyolysis or directly affected by *L. pneumophila*.

The triad of *Legionella* pneumonia, rhabdomyolysis, and AKI is an unusual syndrome that was first reported in 1992 (23) and is associated with high morbidity and mortality rates (up to 40%). We performed a literature search of PubMed using the keywords "*L. pneumophila*" and "rhabdomyolysis" and "acute kidney injury or acute renal injury" There have been 13 published case reports of rhabdomyolysis, renal failure, and Legionnaires' disease. The full text of one case report was unavailable; therefore, 12 case reports were available (Supplementary material). There were 11 men and 1 woman, with an average age of 53 years (range: 26–67). CK levels ranged from 1,103 to 600,000 IU/L, and the serum creatinine ranged from 2.1 to 11.05 mg/dL. All patients received antibiotic therapy with macrolides and/or fluoroquinolones. In total, 7 patients underwent dialysis, and 11 recovered; one female patient and one male patient died. Therefore, an early diagnosis is crucial. We conclude that the timely administration of potent antibacterial drugs and hemodialysis treatment led to the recovery of renal function. All diagnoses were based on antigen or antibody tests. Our patient is the first case of the triad of *Legionella* pneumonia, rhabdomyolysis, and AKI diagnosed using mNGS.

mNGS is an unbiased approach that can theoretically detect all pathogens in clinical samples and is particularly suitable for complicated infectious diseases, including viral, bacterial, fungal, and parasitic diseases (24). mNGS has a significantly better pathogen detection yield than other methods, especially for difficult-to-culture pathogens such as *Legionella* (25–28). Clinicians often find it difficult to distinguish *Legionella* pneumonia from other causes of pneumonia because of the lack of specific clinical manifestations. The gold standard for diagnosis is culture; however, it is now rarely used because it is very time-consuming. Sputum smear microscopy depends on the number of bacteria and the operator's skill. It is difficult to distinguish *Legionella* from other bacteria using microscopic morphology. *Legionella* does not grow on the standard medium used in microbiology laboratories, and a specific medium containing yeast extract and activated charcoal (buffered charcoal yeast extract, BCYE) is required (2). Anti-*Legionella* antibodies in most patients develop only approximately 3 weeks after disease onset, and anti-*Legionella* antibodies are not suitable for patients with severe diseases, such as Legionnaires' disease. In many countries, urinary antigen test is the primary diagnostic technique, although it is poorly sensitive to strains that are non-*L pneumophila* serogroup 1 or other species, including *L. longbeachae* (8). Urine antigens were not available at our hospital. In this case, before the patient came to our hospital, he had been administered piperacillin-tazobactam at a local hospital but still had a fever. The conventional method yielded negative results. After got the sample less than 48 h, the results of mNGS sent to physicians. For fungi, 88,625 reads mapping to *Candida* were detected in BALF. At the same time, no fungi were found in the libraries from the lung biopsy. *Candida* was more common in oral cavity and the upper respiratory tract, so we concluded that some contaminants may be introduced in the process. mNGS detected *Legionella* in both the BALF and lung tissues. Based on his medical history, clinical symptoms, physical signs, results of auxiliary examinations, and mNGS of both BALF and lung tissue, we confirmed the causative pathogen and discontinued meropenem. Thus, this approach also facilitates the use of targeted and efficacious antimicrobial therapies and avoids antibacterial resistance caused by abuse. Indeed, a large number of reads is associated with relative prolonged time and financial costs. Despite the high cost, some severe patients will benefit from the use of mNGS due to timely and targeted treatment. Sometimes, the total cost of patients using multiple pathogen cultures and tests even exceeds mNGS. Meanwhile, with the continuous progress of sequencing technology, the price of mNGS is gradually decreasing.

Legionella is susceptible to erythromycin, clarithromycin, azithromycin, levofloxacin (LEV), and moxifloxacin (MOX) in clinical practice (29). Many antibiotic-resistant clinical isolates of *L. pneumophila* have been identified (30). These isolates were mainly azithromycin-resistant. A meta-analysis concluded that the effectiveness of macrolides or respiratory fluoroquinolones did not reduce mortality among patients with *Legionella* pneumonia (31). Therefore, clinicians should select an antibiotic that is better tolerated and can provide coverage for concomitant infections. Patients with extrapulmonary involvement are often treated with fluoroquinolones (18). These three fluoroquinolones have similar minimum inhibitory concentrations against *L. pneumophila* (30). However, LEV requires dose adjustment according to renal function. Ciprofloxacin should be administered twice daily, which is inconvenient and also requires dose adjustments according to renal function. MOX is an imported drug with proven efficacy and does not require dose adjustment in patients with impaired renal function; therefore, we chose MOX. The current antibiotic treatment strategy entails a 7–10 days course for mild cases and a 21 days course for severe cases, which can be adjusted according to the patient's clinical response. In our case, due to severe illness, we administered MOX for 30 days, including 23 days of intravenous administration and 7 days of oral treatment. Eventually, the patient recovered completely.

This case illustrates the potential for severe rhabdomyolysis in a patient with *Legionella* pneumonia. It is believed that the rapid initiation of precise antimicrobial treatment and early substitution of renal function resulted in good outcomes. mNGS can assist in diagnosing infections caused by difficult-to-culture pathogens early, such as *Legionella*, especially in resource-limited settings where specific assays, such as the urine antigen of *Legionella*, are not accessible.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/ and PRJNA951206.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the participant/patient(s) for the publication of this case report.

Author contributions

RD and YF collected and interpreted the data. RD drafted the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the Sichuan Science and Technology Project (Grant no. 2022YFS0105).

Acknowledgments

We thank Si fen Lu from Precision Medicine Center for helping us with procuring the mNGS data.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1145733/full#supplementary-material

References

1. Gonçalves IG, Simões LC, Simões M. *Legionella pneumophila. Trends Microbiol.* (2021) 29:860–1. doi: 10.1016/j.tim.2021.04.005

2. Viasus D, Gaia V, Manzur-Barbur C, Carratala J. Legionnaires' disease: update on diagnosis and treatment. *Infect Dis Ther.* (2022) 11:973–86. doi: 10.1007/s40121-022-00635-7

3. Simner PJ, Miller S, Carroll KC. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. *Clin Infect Dis.* (2018) 66:778–88. doi: 10.1093/cid/cix881

4. He D, Quan M, Zhong H, Chen Z, Wang X, He F, et al. Emergomyces orientalis emergomycosis diagnosed by metagenomic next-generation sequencing. *Emerg Infect Dis.* (2021) 27:2740–2. doi: 10.3201/eid2710.210769

5. Miao Q, Ma Y, Wang Q, Pan J, Zhang Y, Jin W, et al. Microbiological diagnostic performance of metagenomic next-generation sequencing when applied to clinical practice. *Clin Infect Dis.* (2018) 67:S231–40. doi: 10.1093/cid/ciy693

6. Zhan X-Y, Yang J-L, Sun H, Zhou X, Qian Y-C, Huang K, et al. Presence of viable, clinically relevant *legionella* bacteria in environmental water and soil sources of China. *Microbiol Spectr.* (2022) 10:e0114021. doi: 10.1128/spectrum.01140-21

7. Seegobin K, Maharaj S, Baldeo C, Downes JP, Reddy P. Legionnaires' disease complicated with rhabdomyolysis and acute kidney injury in an AIDS patient. *Case Rep Infect Dis.* (2017) 2017:1–5. doi: 10.1155/2017/8051096

8. Phin N, Parry-Ford F, Harrison T, Stagg HR, Zhang N, Kumar K, et al. Epidemiology and clinical management of Legionnaires' disease. *Lancet Infect Dis.* (2014) 14:1011–21. doi: 10.1016/s1473-3099(14)70713-3

9. Burillo A, Pedro-Botet ML, Bouza E. Microbiology and epidemiology of Legionnaire's disease. *Infect Dis Clin North Am.* (2017) 31:7–27. doi: 10.1016/j.idc.2016.10.002

10. Cunha BA, Burillo A, Bouza E. Legionnaires' disease. *Lancet.* (2016) 387:376–85. doi: 10.1016/S0140-6736(15)60078-2

11. Tokgöz Akyil F, Akyil M, Çoban Agca M, Güngör A, Ozantürk E, Sögüt G, et al. Hyponatremia prolongs hospital stay and hypernatremia better predicts mortality than hyponatremia in hospitalized patients with community-acquired pneumonia. *Tuberk Toraks.* (2019) 67:239–47. doi: 10.5578/tt.68779

12. Schuetz P, Haubitz S, Christ-Crain M, Albrich WC, Zimmerli W, Mueller B. Hyponatremia and anti-diuretic hormone in Legionnaires' disease. *BMC Infect Dis.* (2013) 13:585. doi: 10.1186/1471-2334-13-585

13. Zilberberg MD, Exuzides A, Spalding J, Foreman A, Jones AG, Colby C, et al. Hyponatremia and hospital outcomes among patients with pneumonia: a retrospective cohort study. *BMC Pulm Med.* (2008) 8:16. doi: 10.1186/1471-2466-8-16

14. Beekman R, Duijkers RR, Snijders DD, van der Eerden MM, Kross MM, Boersma WWG. Validating a clinical prediction score for *Legionella*-related community acquired pneumonia. *BMC Infect Dis.* (2022) 22:442. doi: 10.1186/s12879-022-07433-z

15. Miyashita N, Horita N, Higa F, Aoki Y, Kikuchi T, Seki M, et al. Validation of a diagnostic score model for the prediction of *Legionella pneumophila* pneumonia. *J Infect Chemother.* (2019) 25:407–12. doi: 10.1016/j.jiac.2019.03.009

16. Cabral BMI, Edding SN, Portocarrero JP, Lerma EV. Rhabdomyolysis. *Dis Mon.* (2020) 66:101015. doi: 10.1016/j.disamonth.2020.101015

17. Bosch X, Poch E, Grau JM. Rhabdomyolysis and acute kidney injury. *N Engl J Med.* (2009) 361:62–72. doi: 10.1056/NEJMra0801327

18. Kao AS, Herath CJ, Ismail R, Hettiarachchi ME. The triad of Legionnaires' disease, rhabdomyolysis, acute kidney injury: a case report. *Am J Case Rep.* (2022) 23:e936264. doi: 10.12659/AJCR.936264

19. Kellum JA, Romagnani P, Ashuntantang G, Ronco C, Zarbock A, Anders H-J. Acute kidney injury. *Nat Rev Dis Primers.* (2021) 7:52. doi: 10.1038/s41572-021-00284-z

20. Chatzizisis YS, Misirli G, Hatzitolios AI, Giannoglou GD. The syndrome of rhabdomyolysis: complications and treatment. *Eur J Intern Med.* (2008) 19:568–74. doi: 10.1016/j.ejim.2007.06.037

21. Prasanna A, Palmer J, Wang S. Legionnaire's disease presenting with the Legionella triad (pneumonia, rhabdomyolysis, renal failure) and cardiac complications. *Cureus.* (2022) 14:e26056. doi: 10.7759/cureus.26056

22. Soni AJ, Peter A. Established association of legionella with rhabdomyolysis and renal failure: a review of the literature. *Respir Med Case Rep.* (2019) 28:100962. doi: 10.1016/j.rmcr.2019.100962

23. Shah A, Check F, Baskin S, Reyman T, Menard R. Legionnaires' disease and acute renal failure: case report and review. *Clin Infect Dis.* (1992) 14:204–7. doi: 10.1093/clinids/14.1.204

24. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *mBio.* (2015) 6:e01888-15. doi: 10.1128/mBio.01888-15

25. Huang Y, Ma Y, Miao Q, Pan J, Hu B, Gong Y, et al. Arthritis caused by *Legionella micdadei* and *Staphylococcus aureus*: metagenomic next-generation sequencing provides a rapid and accurate access to diagnosis and surveillance. *Ann Transl Med.* (2019) 7:589. doi: 10.21037/atm.2019.09.81

26. Yi H, Fang J, Huang J, Liu B, Qu J, Zhou M. *Legionella pneumophila* as cause of severe community-acquired pneumonia, China. *Emerg Infect Dis.* (2020) 26:160–2. doi: 10.3201/eid2601.190655

27. Wang Y, Dai Y, Lu H, Chang W, Ma F, Wang Z, et al. Case report: metagenomic next-generation sequencing in diagnosis of *Legionella pneumophila* pneumonia in a patient after umbilical cord blood stem cell transplantation. *Front Med.* (2021) 8:643473. doi: 10.3389/fmed.2021.643473

28. Yue R, Wu X, Li T, Chang L, Huang X, Pan L. Early detection of *Legionella pneumophila* and Aspergillus by mNGS in a critically ill patient with *Legionella* pneumonia after extracorporeal membrane oxygenation treatment: case report and literature review. *Front Med.* (2021) 8:686512. doi: 10.3389/fmed.2021.686512

29. Niu S, Zhao L. Metagenomic next-generation sequencing clinches the diagnosis of *Legionella* pneumonia in a patient with acute myeloid leukemia: a case report and literature review. *Front Cell Infect Microbiol.* (2022) 12:924597. doi: 10.3389/fcimb.2022.924597

30. Yang JL, Sun H, Zhou X, Yang M, Zhan XY. Antimicrobial susceptibility profiles and tentative epidemiological cutoff values of *Legionella pneumophila* from environmental water and soil sources in China. *Front Microbiol.* (2022) 13:924709. doi: 10.3389/fmicb.2022.924709

31. Jasper AS, Musuuza JS, Tischendorf JS, Stevens VW, Gamage SD, Osman F, et al. Are fluoroquinolones or macrolides better for treating *Legionella* pneumonia? A systematic review and meta-analysis. *Clin Infect Dis.* (2021) 72:1979–89. doi: 10.1093/cid/ciaa441

frontiers | Frontiers in Public Health

TYPE Methods
PUBLISHED 15 May 2023
DOI 10.3389/fpubh.2023.1184045

Check for updates

OPEN ACCESS

EDITED BY
Ruth Evangeline Timme,
Food and Drug Administration, United States

REVIEWED BY
Jessica C. Chen,
Centers for Disease Control and Prevention (CDC), United States
Eija Trees,
Association of Public Health Laboratories, United States
Michael Feldgarden,
National Institutes of Health (NIH), United States

*CORRESPONDENCE
Katelin V. Gali
✉ katelin.gali@dgs.virginia.gov
Lauren Turner
✉ lauren.turner@dgs.virginia.gov

RECEIVED 10 March 2023
ACCEPTED 06 April 2023
PUBLISHED 15 May 2023

CITATION
Gali KV, St. Jacques RM, Daniels CID, O'Rourke A and Turner L (2023) Surveillance of carbapenem-resistant organisms using next-generation sequencing. *Front. Public Health* 11:1184045. doi: 10.3389/fpubh.2023.1184045

Surveillance of carbapenem-resistant organisms using next-generation sequencing

Katelin V. Gali[1]*, Rachael M. St. Jacques[1], Cheyanne I. D. Daniels[1], Allison O'Rourke[2] and Lauren Turner[1]*

[1]Division of Consolidated Laboratory Services, Department of General Services, Richmond, VA, United States, [2]Division of Clinical Epidemiology, Office of Epidemiology, Virginia Department of Health, Richmond, VA, United States

The genomic data generated from next-generation sequencing (NGS) provides nucleotide-level resolution of bacterial genomes which is critical for disease surveillance and the implementation of prevention strategies to interrupt the spread of antimicrobial resistance (AMR) bacteria. Infection with AMR bacteria, including Gram-negative Carbapenem-Resistant Organisms (CRO), may be acute and recurrent—once they have colonized a patient, they are notoriously difficult to eradicate. Through phylogenetic tools that assess the single nucleotide polymorphisms (SNPs) within a pathogen genome dataset, public health scientists can estimate the genetic identity between isolates. This information is used as an epidemiologic proxy of a putative outbreak. Pathogens with minimal to no differences in SNPs are likely to be the same strain attributable to a common source or transmission between cases. These genomic comparisons enhance public health response by prompting targeted intervention and infection control measures. This methodology overview demonstrates the utility of phenotypic and molecular assays, antimicrobial susceptibility testing (AST), NGS, publicly available genomics databases, and open-source bioinformatics pipelines for a tiered workflow to detect resistance genes and potential clusters of illness. These methods, when used in combination, facilitate a genomic surveillance workflow for detecting potential AMR bacterial outbreaks to inform epidemiologic investigations. Use of this workflow helps to target and focus epidemiologic resources to the cases with the highest likelihood of being related.

KEYWORDS
next-generation sequencing, antimicrobial resistance, outbreak, surveillance, carbapenem-resistant organisms

Introduction

The United States Centers for Disease Control and Prevention (CDC) places Gram-negative carbapenem-resistant organisms (CRO) into the top five most urgent antimicrobial resistance threats in the United States (1). Carbapenem-resistant organisms of public health significance include Enterobacterales order organisms, *Pseudomonas aeruginosa*, and *Acinetobacter baumannii*. Identifying antimicrobial resistance (AMR) genes and disease clusters within the population is essential for preventing and controlling the spread of these pathogens. Next-generation sequencing (NGS) is key to identifying specific resistance genes and their spread through a population. Comparison of pathogens at the nucleotide level using NGS data allows for the determination of relatedness between bacterial isolates. Identifying clusters of closely related bacterial infections by genomic comparison enhances the public health response by enabling targeted intervention and infection control measures.

The Combating Antibiotic Resistant-Bacteria (CARB) initiative began in 2014 and continued with the US National Action Plan for Combating Antimicrobial-Resistant Bacteria, 2020–2025 (2). The initiative spurred the creation of the Antimicrobial Resistance Laboratory Network (ARLN) in 2016 (3). As a CDC ARLN site, Virginia's Division of Consolidated Laboratory Services (DCLS) began receiving CRO submissions in 2017 and implemented testing to identify carbapenemase-producing organisms, antimicrobial susceptibility testing, and PCR resistance gene identification.

In 2019, DCLS began sequencing a subset of Virginia CRO isolates. DCLS utilizes the State Public Health Bioinformatics (StaPH-B) Toolkit (4), a free and open-source python wrapper for various bioinformatics tools and Nextflow-based workflows, to analyze pre-defined AMR datasets. While the workflows are written in the workflow manager language Nextflow, other languages (such as Python, BASH, and JavaScript) and tools are used as well. Each workflow utilizes Docker containers, or compartmentalized tools and their associated dependencies, to produce actionable public health data. Workflows and tools that are hosted in the StaPH-B Toolkit are developed by a U.S. public health laboratory consortium (5) and are subjected to rigorous validation and verification processes. Each of the discussed bioinformatics tools included herein (except for National Center for Biotechnology Information (NCBI) Pathogen Detection) are included in the Toolkit.

Public health laboratories receiving CDC funding for CRO sequencing are required to submit sequences to NCBI Pathogen Detection (6, 7). One advantage of submission to Pathogen Detection is for broad swath surveillance for potential genetically related isolates among all reads submitted to NCBI under organism-specific umbrella BioProjects surveilled by Pathogen Detection. Adopting NGS methods for detection of clusters of AMR bacterial isolates, as well as identification of the underlying resistance mechanisms harbored, varies substantially between laboratories. While ongoing development within public health laboratories for more efficient and actionable utilization of NGS data continues, genomic comparison has proven useful in detecting and controlling outbreaks of AMR infections (8, 9). By harnessing the aforementioned services and bioinformatics software, a tiered workflow for surveillance of resistance genes and identification of potential disease clusters of these pathogens was piloted and is proposed for consideration by the broader public health community.

Materials and methods

Microbiology methods

Carbapenem resistance screening for organisms that have acquired a carbapenemase-producing gene begins by testing bacterial cultures for carbapenemase enzyme production using the modified Carbapenem Inactivation Method (mCIM) (10). All mCIM-positive isolates receive PCR testing using the Streck^{TM} ARM-D β-lactamase PCR kit and antimicrobial susceptibility testing using the Sensititre^{TM} Gram Negative MIC GN7F Plate (ThermoFisher Scientific, Waltham, Massachusetts). The Streck PCR assay detects the presence of the five most common carbapenemase genes (KPC, NDM, VIM, IMP, and OXA-48). Further genomic characterization using next-generation sequencing is performed on isolates meeting one of the following CDC ARLN criteria: (i) Enterobacterales PCR-positive for any carbapenemase gene other than, or in addition to KPC, due to the high prevalence of KPC-positive isolates in Virginia, (ii) *Pseudomonas aeruginosa* and *Acinetobacter baumannii* isolates that are PCR positive for any carbapenemase gene, including KPC, due to the low KPC-positivity for these organisms in Virginia, (iii) Enterobacterales, *P. aeruginosa*, or *A. baumannii* isolates with resistance or non-susceptibility to all drugs in the Sensititre panel and the submitting facility's antimicrobial susceptibility testing panel. (iv) organisms that are PCR-positive for two or more carbapenemase genes; (v) mCIM-positive and PCR-negative cultures which may harbor a novel resistance mechanism (11). Of these criteria, novel resistance is the highest priority for identification of emerging resistance factors.

Extraction and sequencing methods

Manual DNA extraction

Carbapenem-resistant genomic DNA is extracted using QIAGEN QIAamp DNA Mini Kit (Qiagen, Aarhus, Denmark) from isolated bacterial colonies grown on Tryptic Soy Agar (TSA) with 5% sheep blood agar (Remel, Lenexa, Kansas) for 18–24 h at 33–37°C. The following modifications were implemented to the QIAGEN QIAamp DNA Mini Kit (12) method to obtain optimal total DNA for short-read sequencing. Cell lysis is performed in a biosafety cabinet to render the isolate no longer infectious. A loopful of isolated bacterial colonies from the TSA with 5% sheep blood agar plate is added into a labeled 1.5 mL safe-lock tube with 180 μL of ATL buffer, vortexed and pulse-centrifuged. Proteinase K (20 μL) is added to the sample and incubated at 56°C ± 1 for 1–3 h with vortexing every 20 mins. Immediately after incubation, 4 μL of RNase A (Qiagen, Aarhus, Denmark) is added to the sample, held at room temperature for 3–5 mins, followed by 200 μL of AL buffer and 200 μL of 100% ethanol (Pharmco, Brookfield, Connecticut). For quality assurance, a blank sample, or no template control (NTC) is carried throughout the

Abbreviations: AMR, Antimicrobial Resistance; ARLN, Antimicrobial Resistance Laboratory Network; AST, Antimicrobial Susceptibility Testing; CARB, Combating Antibiotic Resistant Bacteria; CDC, Centers for Disease Control and Prevention; CRO, Carbapenem Resistant Organism; DCLS, Division of Consolidated Laboratory Services; GAMMA, Gene Allele Mutation Microbial Assessment; HAI, Healthcare-Associated Infection; IMP, Imipenemase; KPC, *Klebsiella pneumoniae* carbapenemases; Mbp, megabase pair; mCIM, modified Carbapenem Inactivation Method; MIC, Minimal Inhibitory Concentration; NaOH, Sodium hydroxide; NCBI, National Center for Biotechnology Information; NDM, New Delhi Metallo-Beta-Lactamase; NGS, Next-generation sequencing; OXA, Oxacillinase; PCR, Polymerase Chain Reaction; SNPs, Single Nucleotide Polymorphisms; StaPH-B, State Public Health Bioinformatics; VDH, Virginia Department of Health; VIM, Verona Intergron Encoded Metallo-Beta-Lactamase; WGS, Whole Genome Sequencing.

extraction and sequencing procedures to assess contamination or other quality errors in testing.

Following cell lysis, samples are safely manipulated on the bench top for DNA cleanup. The entire cell lysis volume is transferred to a spin column placed in a 2 mL collection tube and centrifuged at 6,000 x g for 1 min to bind genomic DNA (gDNA) to the spin column's silica membrane. Then, following the manufacturer's protocol, the spin column is washed twice using 500 μL of AW1 and AW2 buffers at 6,000 x g for 1 min and 20,000 x g for 3 mins, respectively, and then eluted into a clean tube using 100 μL of 10 mM Tris-HCl, pH 8 (Fisher Scientific, Hampton, New Hampshire). DNA quantification post extraction is measured with the Qubit dsDNA Broad Range Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts) on a Qubit fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts) to remove any samples with suboptimal concentration (≤5 ng/μL) from further testing.

Whole genome sequencing (WGS)

The number of samples per sequencing run is determined by the 500-cycle MiSeq Reagent Kit v2 (Illumina, San Diego, CA), which has a maximum output of 8.5 Gb. For optimal run quality, the total genome load for a 500-cycle cartridge is limited to 100 megabase pairs (Mbps), equivalent to up to 20 cultures with 5 Mbp genomes (13). A diverse run composition of bacterial species is selected for library preparation. However, GC-rich content organisms, such as *P. aeruginosa*, are limited to 4 to 6 samples per run to avoid bias in sample read coverage (14).

WGS of bacterial isolates includes six components: library preparation, quantification, optional fragment analysis, normalization, denaturation, and loading (15). Samples are prepared for WGS using the Illumina DNA Prep kit (Illumina, San Diego, CA) with an average of 100–500 ng of input gDNA per sample for a total volume of 30 μL. The library clean-up procedure has been modified to utilize 40. 8 μL SPB/IPB and 44.2 μl H_2O per sample, to capture longer DNA fragments (13). Quantification and fragment analysis is recommended at the end of preparation to evaluate the quality of individual DNA and pooled libraries. The blank (NTC control) is not loaded in the final pool but is assessed for quality using the Qubit fluorometer (see below).

Individual DNA and pooled libraries are quantified using the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts). Libraries prepared using the Illumina DNA Prep method have an average quantification value of 10 ng/μL; however, the quantification value can vary. The allowable quantification values for library blanks are ≤ 0.1 ng/μL for Qubit 2.0 and "out of range" for Qubit 3 and 4. Fragment analysis is completed using the Agilent D5000 ScreenTape kit and 4200 TapeStation System. Average fragment sizes are obtained using the region view, usually 800–1,000 bp.

Samples can be normalized individually; however, this procedure uses the pool normalization method. This method takes the pool concentration and average fragment size to calculate the molarity of DNA from the pooled libraries, molarity (nM) = [(Pool concentration ng/μL) / (660 g/mol x fragment size bp)] x 10^6. The preferred starting library concentration for denaturation and loading is 4 nM. The formula M1V1 = M2V2 calculates the amount of pooled library required to achieve 50 μL of a 4 nM pool (200 / molarity). The volume of the pool required is then subtracted from 50 μL to determine the volume of diluent. The 4 nM pool is denatured with 0.2 N NaOH and denaturation is halted, and the 4 nM pool is further diluted using 990 μL of HTl. At this step, the denatured pool has a concentration of 20 pM and will be diluted for optimal clustering. The formula C1V1 = C2V2 is applied to calculate the amount of denatured pool required to achieve a final loading concentration of 15 pM, (20 pM) V1 = (15 pM) (1,000 mL).

DNA sequencing is performed on the Illumina MiSeq Sequencing System using the 500 cycle v2 (2 x 251) base pair sequencing chemistry. A PhiX Control v3 Library (Illumina, San Diego, CA) is helpful for troubleshooting issues with cluster density related to library preparation. The PhiX solution is denatured and diluted to match the pooled library at 15 pM and is spiked into the final pool at 1%. The denatured DNA/PhiX library pool is heated at 96°C ± 1 for 2 mins and submerged in ice for 5 mins before transferring 600 μL into the 500-cycle MiSeq cartridge.

Sample sheets are built on Local Run Manager (16). The cartridge and buffers are loaded into the instrument. MiSeq Control Software is used to start the WGS run which requires a BaseSpace account to access sequencing data for analysis. Prior to starting the run, the MiSeq will do a system check to verify the run parameters, reagent radio-frequency identification (RFID), available disk space, and internet access. Following the pre-run check, the run is started and takes ∼36 h.

Post-run metrics are reviewed to assess the overall run quality. If critical run metrics pass (see Table 1), the run is accepted for initial bioinformatics analyses. Runs with quality metrics below the expected results are comprehensively reviewed for troubleshooting purposes and reloaded from library preparation. Run performance can vary depending on run composition, library preparation, and instrument errors; however, the Illumina Sequencing Analysis Viewer can be used to investigate possible solutions (17).

Bioinformatics methods

Machine configuration

Bioinformatics analyses were performed using Amazon Web Service Elastic Cloud Computing (AWS EC2) environments with base Ubuntu 18.04 Bionic image virtual machines (VMs) with a T2.2xlarge image (8 vCPUs, 32 GB of RAM).

Tredegar

The DCLS-developed and validated Tredegar pipeline was used to analyze short-read Illumina data for quality and taxonomic label verification of WGS data (18). Once sequencing run data is pushed from the MiSeq instruments to BaseSpace, the data is pulled from the cloud and analyzed on the VMs for quality control.

The following command was used for each analysis:

```
$ staphb-wf tredegar -o
<output_directory> <path/to/reads>
```

After the data is pulled from BaseSpace, Tredegar is utilized to calculate the average read quality for both forward and reverse reads. Minimum data acceptability criteria include (i) *fastq* Q scores

TABLE 1 Post run quality metrics.

Quality metrics	Cluster passing filter	Q30 (%)	Q30 R1/R2 (%)	Cluster density (K/mm^2)	Estimated yield (Gb)	Aligned PhiX (%)	PhiX error rate (%)
Expected results	≥80	75	N/A	600–1,200	N/A	0.88–1.85	0.92–1.45

≥ 30 for both the forward and reverse reads (r1_q and r2_q, respectively), (ii) an estimated genome length (est_genome_length) within 0.5 Mbps of the expected genome size as determined on the NCBI Genome browser (19), (iii) an estimated coverage (est_cvg) ≥ 40x (the total number of bases generated for the run divided by the assembly length estimated from the *de novo* Shovill assembly (20), (iv) assembled contig (number_contigs) <200, and (v) the species prediction (species_prediction) by MASH (21) must match the organism determined by MALDI-TOF Mass Spectrometry. Deviation from these metrics may point to contamination, sample switching, or sequencing malfunction.

Tredegar analyses are reported to the sequencing scientist via custom-designed CSV files (Table 2). Isolates with quality metrics failing to meet the above criteria are rejected and excluded from further bioinformatics analysis. Sequences meeting quality metrics are submitted to NCBI Pathogen Detection.

NCBI pathogen detection

Illumina sequencing reads and minimum isolate metadata (excluding patient identifiable information), are submitted to the NCBI Sequence Read Archive and CDC HAI-Seq Umbrella Project, Gram Negative Bacteria BioProject PRJNA288601 with a unique sample identification number assigned by the sequencing laboratory for sample anonymity (6). Submission to an Umbrella BioProject linked to NCBI Pathogen Detection prompts the automatic analysis of reads for integration into the Pathogen Detection Project (7, 22). In Pathogen Detection, there are two different clustering pipelines in operation. For organisms which have a whole genome multiple locus sequence type (wgMLST) scheme available, a reference wgMLST scheme is used to identify the loci and alleles in each assembled genome, and then a 25-allele cut-off is applied to identify potential cluster related isolates. The second process for organisms with less than 1,000 isolates on Pathogen Detection, or for which there is not a wgMLST scheme utilizes k-mer distances to first cluster related isolates, and then a first pass SNP analysis. Clusters are created using 50-SNP single-linkage clustering. Once clusters are created by the wgMLST or K-mer process, a reference is selected within each, assemblies are aligned against the reference, SNPs are called, and phylogenetic trees inferred. The sizes of clusters may vary from two isolates to thousands, and for each organism group isolates which do not fall within the cluster detection criteria are omitted (23). The cluster analysis process automatically starts once daily for each organism, if new data are submitted.

Pathogen Detection provides AMR gene prediction for all submitted isolates in addition to SNP distances and phylogenetic trees for clustered isolates (22, 24). An email notification alert was built to alert analysts when a submitted isolate is added into a SNP cluster on NCBI. In the DCLS surveillance workflow, NCBI provides the initial phylogenetic and cluster analysis.

Hickory

For analysis of a pre-defined organism dataset, the DCLS-developed Hickory (25) bioinformatics pipeline was used to determine the most appropriate reference genome within the Illumina short-read dataset via MASH (21). The Hickory pipeline takes in fastq files and generates assemblies from the data. Once the fasta files have been generated, binary sketches of the fasta files are drawn within an individual directory using MASH. The fasta file sketch, or genome sketch, with the least MASH distance from the other fasta file sketches in the directory is selected as the most appropriate reference genome. The selection of a reference genome with the least distance from the dataset is important because it increases the number of nucleotide positions available for comparative genomics, and therefore, inferences made about genomic similarity or dissimilarity of a dataset. Hickory provides the reference-free FASTA assembly file of the appropriate reference genome for each dataset analyzed. This FASTA file is then used as the reference genome during Dryad analysis. Hickory ensures a closely related reference is used for comparative genomic analysis so that the maximum number of positions can be queried.

After separating the read data by species, the following command was used for each analysis:

```
  $ staphb-wf hickory -o
<output_directory> <path/to/reads>
```

Dryad

Isolates that pass Tredegar metrics are analyzed by Dryad, a bioinformatics tool developed by the Wisconsin State Public Health Labs and validated by DCLS (26). Dryad utilizes the CFSAN SNP pipeline to determine the SNP distance between closely related samples (27). Potential AMR determinants are identified via AMRFinder Plus (22, 24).

After separating the read data by species, the following command was used for each analysis:

```
  $ staphb-wf dryad main -cg -s -r
<reference.fasta> -o <output_dir> /
  -report <reads>
```

Dryad analyses produce SNP-distance heatmaps, phylogenetics trees, and if selected during the analysis initiation, a list of AMR gene predictions. Isolates that are ≤ 11 SNPs apart are considered "putative" outbreak clusters; bioinformaticians rely on epidemiologists and their gathered evidence to determine if an isolate is truly related. Isolates that are between 12 and 15 SNPs apart can still be considered related with enough supporting epidemiological evidence. Carbapenem-resistant isolates can be

TABLE 2 Example of Tredegar results with passing quality metrics.

Sample	rq_1	r2_q	est_genome_length	est_cvg	number_contigs	species_prediction	subspecies_ prediction
2022EP-00093	35.26	31.45	5524279	81.33	74	Klebsiella_pneumoniae	NA
2022EP-00091	37.09	34.98	3807676	85.89	51	Acinetobacter_baumannii	NA
2022EP-00092	35.04	31.55	5398118	59.17	47	Serratia_marcescens	NA
2021EP-00104	36.91	35.32	3871668	145.61	93	Acinetobacter_baumannii	NA
2021EP-00106	36.7	34.64	3937667	105.59	100	Acinetobacter_baumannii	NA
2022EP-00007	35.69	32.16	5313287	102.98	147	Escherichia_coli	O102:H6

considered related at a larger SNP range than other isolates; isolates that are between 12 and 30 SNPs apart may be determined to be related with epidemiological support.

Individual introduction cases are determined by the number of SNPs separating the isolates in an outbreak dataset. For example, Figure 1 shows isolates VA7, VA6, and VA8 are between 1 and 4 SNPs apart from one another. Isolates VA1 and VA2 are 0 SNPs apart from one another. This shows two putative clusters in the dataset; Group A, composed of VA7, VA6, and VA8, and Group B, composed of VA1 and VA2. These results indicate the presence of two putative outbreak clusters, or two separate introductions.

GAMMA

GAMMA (28), Gene Allele Mutation Microbial Assessment, is a CDC-developed bioinformatics software tool designed to analyze FASTA files to identify protein coding regions of interest. Currently, DCLS is utilizing a CDC provided custom database to elucidate hyper-virulence genes (*peg-344*, *iroB*, *iucA*, *$_{p}$rmpA*, and *$_{p}$rmpA2*) from sequencing assembly. GAMMA uses a Conda environment during routine analyses. GAMMA result TSV files are passed to the requesting scientists for epidemiology-report generation. Hypervirulence genes identified by GAMMA are submitted to the CDC ARLN branch.

The following command was used for each analysis:

```
$ GAMMA.py fasta_file
custom_db.fasta output_dir
```

Results

NCBI cluster surveillance

In November 2021, DCLS began piloting a program using NCBI Pathogen Detection in a tiered surveillance method. Table 3 demonstrates the value in using NCBI Pathogen Detection as the primary step in the surveillance method. Of the 381 isolates sequenced from 2019 when CRO sequencing began until May 2022, 104 cluster notifications would have been received that include Virginia isolates. After removing clusters that only included multiple isolates from the same patient, 91 clusters would have prompted further investigation. Many of these clusters carry over from 1 year to the next due to the long-term colonization of patients and environments with resistant organisms.

Once a cluster is identified, scientists review the notification email from NCBI. The DCLS criteria for potential outbreak surveillance are more stringent (≤ 11 SNPs) compared to NCBI (≤50 SNPs). As mentioned previously, isolates between 12 and 30 SNPs may be included if there is epidemiologic evidence. Scientists will identify cluster isolates within 11 SNPs of each other and verify there are at least three Virginia isolates in the cluster (per epidemiologist request). Bioinformaticians use Hickory and Dryad for a more thorough investigation of the identified cluster. Dryad assesses and confirms the SNP distances between the clustered isolates using a within-dataset reference genome determined by Hickory. Both Dryad and Hickory utilize well established, open source, peer reviewed bioinformatics tools and have been validated through a rigorous state validation process. Once the Dryad pipeline confirms the SNP distance and AMR gene prediction results from NCBI Pathogen Detection, DCLS scientists build a surveillance report based on the combined wet-lab and bioinformatics results to communicate the findings to the epidemiologist.

The surveillance report includes the SNP matrix, resistance gene predictions confirmed by PCR (ex: NDM, VIM, KPC, OXA-48, or IMP), patient identification, and AST results. Since March 2022, the implementation of the NGS surveillance process has resulted in 30 communications of potential outbreak clusters. Five of these were for Enterobacterales and *Pseudomonas aeruginosa* which are provided in a surveillance report to epidemiologists at the Virginia Department of Health (VDH). The other communications were for *Acinetobacter baumannii* isolates which are of secondary priority to VDH epidemiologists and per request, cluster information is e-mailed to the epidemiologist. All NGS result reports include a disclaimer stating results are not for clinical diagnosis or patient management but are for epidemiologic purposes only. NGS results are communicated only to the health department epidemiologists.

While Dryad is a useful tool for analyzing individual outbreak datasets, the process is reliant on scientists submitting requests for known isolates. By utilizing the NCBI cluster detection pipeline, DCLS has begun to identify and analyze outbreaks both within-state and nationally. Since NCBI Pathogen Detection includes submissions from other laboratories, including other public health laboratories, cluster notifications can include DCLS isolates, and closely related isolates sequenced at other public health laboratories. Informing epidemiologists of these potential links to out-of-state isolates assists in determining possible sources of infection or enabling multi-state investigations.

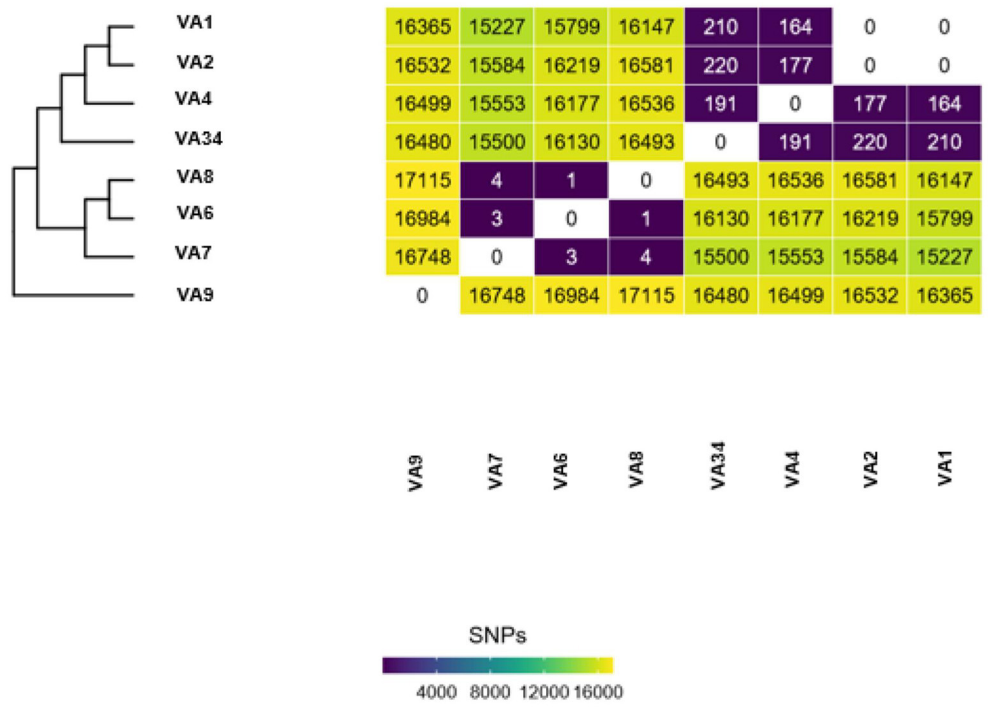


FIGURE 1
Example output of Dryad data. Each Dryad analysis produces a SNP-distance heatmap and phylogenetic tree.

TABLE 3 Isolates sequenced and NCBI clusters identified.

	Isolates sequenced	NCBI clusters	Adjusted-removed same patient clusters	Adjusted-removed clusters from other years
2019	148	39	32	32
2020	102	30	26	17
2021	119	30	28	17
2022	12	5	5	2
Total	**381**	**104**	**91**	**68**

For example, during an outbreak investigation of *Proteus mirabilis* isolates in Virginia for a local health department, the NCBI Pathogen Detection pipeline identified several other isolates sequenced by the Mid-Atlantic ARLN regional laboratory. Communication with the regional laboratory found that the isolates were colonization screening specimens sent from Virginia to the regional laboratory since DCLS does not currently perform colonization screening. Use of NGS for surveillance increased the number of isolates potentially related to this outbreak from 3 to 10 spanning a much longer period than originally investigated (Figure 2, *Proteus mirabilis* Cluster).

Figure 2 shows all the clusters meeting surveillance notification criteria during a pilot of the tiered surveillance method DCLS performed from November 2021 to April 2022. NGS surveillance provided many previously unidentified clusters and isolates (Shown in blue in Figure 2, Surveillance WGS link). As has been demonstrated by PulseNet for foodborne diseases, the ability to link cases of related infections using NGS is a powerful epidemiologic tool (29, 30). Surveillance and identification of related antimicrobial-resistant isolates provides an increased ability to respond and prevent the spread of this serious public health threat.

Dryad/NCBI SNP discrepancies

Though rare, differences can occur in the SNP calls from the different pipelines. Several factors can cause discrepancies between these tools. Dryad is an open-source bioinformatics tool which lists each tool used and its version. NCBI Cluster Detection uses a suite of alternative tools curated by NCBI to perform assembly, genome annotation, antibiotic resistance determination and genome clustering (31, 32). Slight differences in tool heuristics and their parameters can lead to variations in SNP distances (33, 34). Reference genome selection can affect the SNP distances because the reference genome is the sequence to which all other cluster isolates are compared and if a more distant reference genome is used, there is a risk of losing genomic comparability for regions absent in the reference (35). The NCBI reference genome selection method chooses an in-group reference genome with the longest read from an initial dataset, which is often larger since it includes sequences from other NCBI submitters (31). Hickory selects the reference genome from within the user-defined dataset

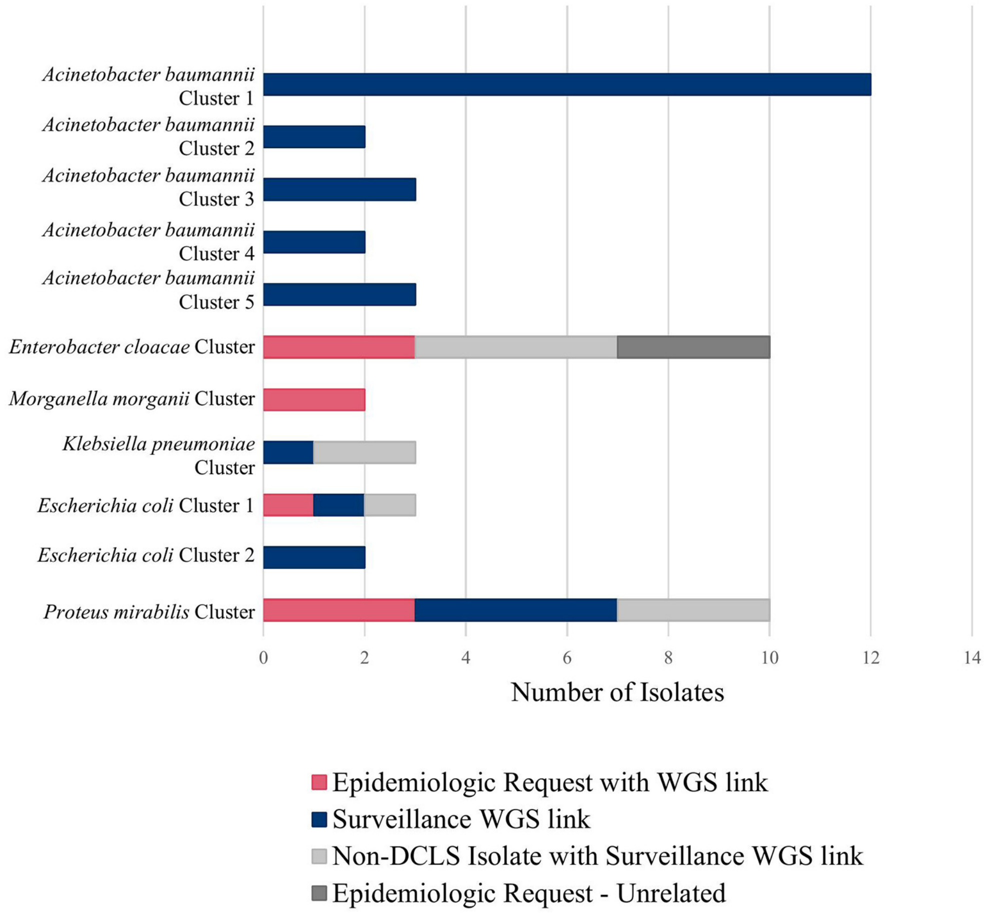


FIGURE 2
Clusters detected using tiered surveillance method (November 2021–April 2022).

based on MASH distance (21). The user-defined dataset typically consists of isolates only sequenced and under suspicion of being outbreak-associated. Theoretically, there is an increased likelihood of identifying a greater number of SNPs because within dataset selected genomes should have a greater genetic identity. Masking portions of the genome sequence can also lead to differences in SNP distances. Some tools mask repetitive genome regions before SNP analysis, potentially altering downstream data. NCBI utilizes masking, while Dryad does not. Computing resources can also influence downstream analysis results (36). Masking and reference genome selection are the most likely causes of the significant discrepancies shown in Table 4.

TABLE 4 Dryad and NCBI discrepancies.

Cluster	Isolate #	Dryad	NCBI
Proteus mirabilis	2022EP-00001 & 7 isolate cluster	16-35 SNPs	4-10 SNPs
Acinetobacter baumannii	2021EP-00086 & 2021EP-00090	64-75 SNPs	7-14 SNPs
Enterobacter cloacae	2019EP-00005 & 2019EP-00121	479 SNPs	19 SNPs
Klebsiella pneumoniae	2022EP-00149 & 2022EP-00173	292 SNPs	18 SNPs

NDM-19/NDM-7 cluster

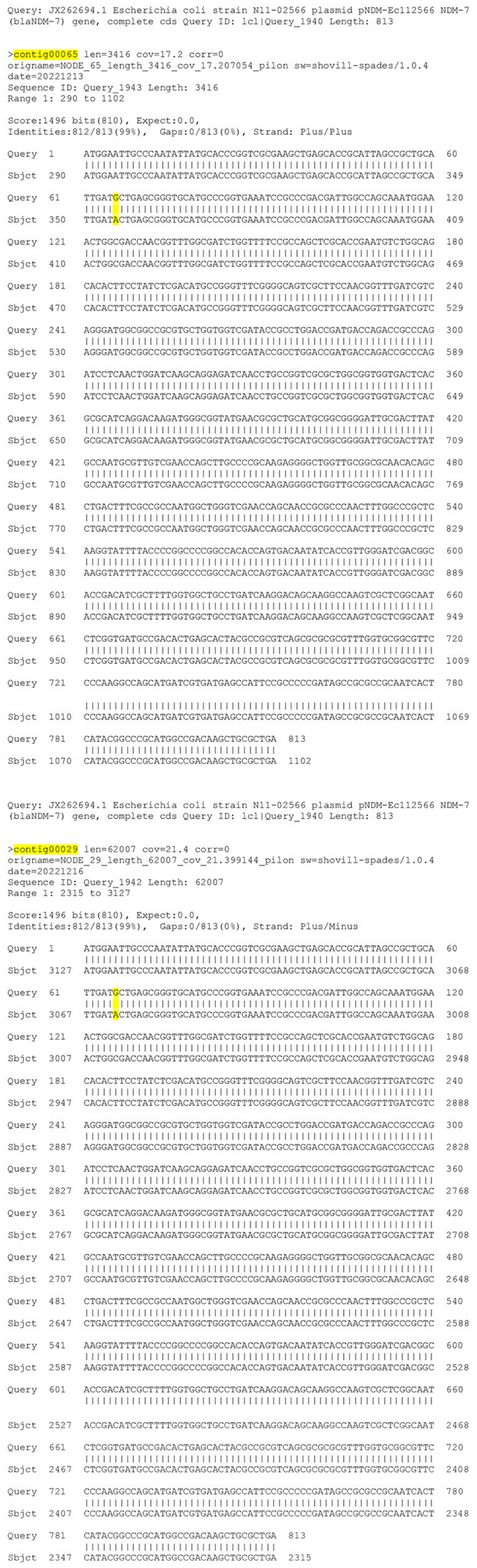
Another benefit to running multiple analysis tools is using repeat results as a check to ensure the result report includes all the resistance genes present in the genome. On rare occasions, one AMR analysis tool doesn't report a gene found by another AMR prediction tool, and further analysis is required to verify the results. One example of when running two analysis tools proved beneficial was with an NDM *Klebsiella pneumoniae* outbreak investigation. On NCBI Pathogen Detection, AMR prediction of one isolate (2022EP-00101) had an NDM-19 gene, and the other isolate (2022EP-00107) had an NDM-7 gene. Dryad analysis results lacked the NDM-7 on 2022EP-00107 in the DCLS AMR report. Both isolates had positive NDM results from the Streck ARM-D, β-lactamase PCR kit further verifying the NCBI results. Re-sequencing and repeat Dryad analysis produced the same results. Combining the reads from both sequencing runs provided more depth and coverage, and Dryad analysis of the combined assembly identified the NDM-7 gene on the AMR prediction profile. Using more than one tool proved significant because results from one bioinformatics tool showed a gap in the results identified by the second tool, and further supported the overall SNP comparison indicating isolate genetic identity.

AMR genes reporting

AMR gene predictions can produce a long list of resistance genes. Determining which genes to report to the epidemiologist has been a significant challenge. Including a list of all the genes identified can be overwhelming and not always informative or helpful since epidemiologists already have the antimicrobial susceptibility testing (AST) results, and prediction of a gene is not equivalent to expression. At DCLS, reporting AMR genes is based on the significance of the gene within the isolate or outbreak cluster, as determined by relevance to other phenotypic testing by mCIM, AST and PCR. A gene and allele number are always provided for carbapenemase genes since these genes are of interest to the CDC.

For example, knowing an isolate or outbreak cluster carries the NDM-5 gene will provide specific information on which resistance gene is responsible for the carbapenem resistance. Allele identification also enables tracking of the frequency of individual alleles within a geographic area. If an allele is unknown, meaning the beta-lactamase (*bla*) gene is returned un-numbered by the AMR prediction method (ex: NDM-5 vs. NDM), further analysis is necessary to verify the presence of a unique allele and identify the responsible genome mutation (24). Using an alignment tool to compare the un-numbered gene sequence to the closest neighbor allows for identification of the nucleotide differences between the genes. Requesting an allele number for these un-numbered alleles is done through NCBI (37). Recently DCLS identified an un-numbered NDM allele that was a mutation of an NDM-7 in a *K. pneumoniae*. This gene had an M22I amino acid change due to a point mutation (Figure 3). The novel gene was named NDM-50 by NCBI. Surveillance detected two other isolates with this gene in Virginia over the next 2 months. Genome alignment can also be used to verify gene mutations of numbered alleles in closely related isolates as in the NDM-19/NDM-7 cluster above.

Currently, NGS result reports to epidemiologists include only genes indicated by PCR-detection. If an antimicrobial resistance gene other than a carbapenemase gene is predicted in one outbreak isolate, but absent in another, that could explain the difference in susceptibility results of a specific drug in the isolate's AST profiles, then a comment is added to the surveillance report shared with the health department. The report states that the difference in the AST profile is most likely due to the presence or absence of a resistance gene without naming the specific gene. The report may also include other novel genes causing carbapenem resistance identified in the organism of interest. Each report includes a disclaimer stating results are for epidemiologic purposes only and not for clinical diagnosis or patient management.



```
Query: JX262694.1 Escherichia coli strain N11-02566 plasmid pNDM-Ec112566 NDM-7
(blaNDM-7) gene, complete cds Query ID: lcl|Query_1940 Length: 813


>contig00065 len=3416 cov=17.2 corr=0
origname=NODE_65_length_3416_cov_17.207054_pilon sw=shovill-spades/1.0.4
date=20221213
Sequence ID: Query_1943 Length: 3416
Range 1: 290 to 1102

Score:1496 bits(810), Expect:0.0,
Identities:812/813(99%),  Gaps:0/813(0%), Strand: Plus/Plus

Query  1     ATGGAATTGCCCAATATTATGCACCCGGTCGCGAAGCTGAGCACCGCATTAGCCGCTGCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  290   ATGGAATTGCCCAATATTATGCACCCGGTCGCGAAGCTGAGCACCGCATTAGCCGCTGCA  349

Query  61    TTGATGCTGAGCGGGTGCATGCCCGGTGAAATCCGCCCGACGATTGGCCAGCAAATGGAA  120
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  350   TTGATACTGAGCGGGTGCATGCCCGGTGAAATCCGCCCGACGATTGGCCAGCAAATGGAA  409

Query  121   ACTGGCGACCAACGGTTTGGCGATCTGGTTTTCCGCCAGCTCGCACCGAATGTCTGGCAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  410   ACTGGCGACCAACGGTTTGGCGATCTGGTTTTCCGCCAGCTCGCACCGAATGTCTGGCAG  469

Query  181   CACACTTCCTATCTCGACATGCCGGGTTTCGGGGCAGTCGCTTCCAACGGTTTGATCGTC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  470   CACACTTCCTATCTCGACATGCCGGGTTTCGGGGCAGTCGCTTCCAACGGTTTGATCGTC  529

Query  241   AGGGATGGCGGCCGCGTGCTGGTGGTCGATACCGCCTGGACCGATGACCAGACCGCCCAG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  530   AGGGATGGCGGCCGCGTGCTGGTGGTCGATACCGCCTGGACCGATGACCAGACCGCCCAG  589

Query  301   ATCCTCAACTGGATCAAGCAGGAGATCAACCTGCCGGTCGCGCTGGCGGTGGTGACTCAC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  590   ATCCTCAACTGGATCAAGCAGGAGATCAACCTGCCGGTCGCGCTGGCGGTGGTGACTCAC  649

Query  361   GCGCATCAGGACAAGATGGGCGGTATGAACGCGCTGCATGCGGCGGGGATTGCGACTTAT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  650   GCGCATCAGGACAAGATGGGCGGTATGAACGCGCTGCATGCGGCGGGGATTGCGACTTAT  709

Query  421   GCCAATGCGTTGTCGAACCAGCTTGCCCCGCAAGAGGGGCTGGTTGCGGCGCAACACAGC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  710   GCCAATGCGTTGTCGAACCAGCTTGCCCCGCAAGAGGGGCTGGTTGCGGCGCAACACAGC  769

Query  481   CTGACTTTCGCCGCCAATGGCTGGGTCGAACCAGCAACCGCGCCCAACTTTGGCCCGCTC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  770   CTGACTTTCGCCGCCAATGGCTGGGTCGAACCAGCAACCGCGCCCAACTTTGGCCCGCTC  829

Query  541   AAGGTATTTTACCCCGGCCCCGGCCACACCAGTGACAATATCACCGTTGGGATCGACGGC  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  830   AAGGTATTTTACCCCGGCCCCGGCCACACCAGTGACAATATCACCGTTGGGATCGACGGC  889

Query  601   ACCGACATCGCTTTTGGTGGCTGCCTGATCAAGGACAGCAAGGCCAAGTCGCTCGGCAAT  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  890   ACCGACATCGCTTTTGGTGGCTGCCTGATCAAGGACAGCAAGGCCAAGTCGCTCGGCAAT  949

Query  661   CTCGGTGATGCCGACACTGAGCACTACGCCGCGTCAGCGCGCGCGTTTGGTGCGGCGTTC  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  950   CTCGGTGATGCCGACACTGAGCACTACGCCGCGTCAGCGCGCGCGTTTGGTGCGGCGTTC  1009

Query  721   CCCAAGGCCAGCATGATCGTGATGAGCCATTCCGCCCCCGATAGCCGCGCCGCAATCACT  780

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1010  CCCAAGGCCAGCATGATCGTGATGAGCCATTCCGCCCCCGATAGCCGCGCCGCAATCACT  1069

Query  781   CATACGGCCCGCATGGCCGACAAGCTGCGCTGA  813
             |||||||||||||||||||||||||||||||||
Sbjct  1070  CATACGGCCCGCATGGCCGACAAGCTGCGCTGA  1102


Query: JX262694.1 Escherichia coli strain N11-02566 plasmid pNDM-Ec112566 NDM-7
(blaNDM-7) gene, complete cds Query ID: lcl|Query_1940 Length: 813


>contig00029 len=62007 cov=21.4 corr=0
origname=NODE_29_length_62007_cov_21.399144_pilon sw=shovill-spades/1.0.4
date=20221216
Sequence ID: Query_1942 Length: 62007
Range 1: 2315 to 3127

Score:1496 bits(810), Expect:0.0,
Identities:812/813(99%),  Gaps:0/813(0%), Strand: Plus/Minus

Query  1     ATGGAATTGCCCAATATTATGCACCCGGTCGCGAAGCTGAGCACCGCATTAGCCGCTGCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3127  ATGGAATTGCCCAATATTATGCACCCGGTCGCGAAGCTGAGCACCGCATTAGCCGCTGCA  3068

Query  61    TTGATGCTGAGCGGGTGCATGCCCGGTGAAATCCGCCCGACGATTGGCCAGCAAATGGAA  120
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3067  TTGATACTGAGCGGGTGCATGCCCGGTGAAATCCGCCCGACGATTGGCCAGCAAATGGAA  3008

Query  121   ACTGGCGACCAACGGTTTGGCGATCTGGTTTTCCGCCAGCTCGCACCGAATGTCTGGCAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3007  ACTGGCGACCAACGGTTTGGCGATCTGGTTTTCCGCCAGCTCGCACCGAATGTCTGGCAG  2948

Query  181   CACACTTCCTATCTCGACATGCCGGGTTTCGGGGCAGTCGCTTCCAACGGTTTGATCGTC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2947  CACACTTCCTATCTCGACATGCCGGGTTTCGGGGCAGTCGCTTCCAACGGTTTGATCGTC  2888

Query  241   AGGGATGGCGGCCGCGTGCTGGTGGTCGATACCGCCTGGACCGATGACCAGACCGCCCAG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2887  AGGGATGGCGGCCGCGTGCTGGTGGTCGATACCGCCTGGACCGATGACCAGACCGCCCAG  2828

Query  301   ATCCTCAACTGGATCAAGCAGGAGATCAACCTGCCGGTCGCGCTGGCGGTGGTGACTCAC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2827  ATCCTCAACTGGATCAAGCAGGAGATCAACCTGCCGGTCGCGCTGGCGGTGGTGACTCAC  2768

Query  361   GCGCATCAGGACAAGATGGGCGGTATGAACGCGCTGCATGCGGCGGGGATTGCGACTTAT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2767  GCGCATCAGGACAAGATGGGCGGTATGAACGCGCTGCATGCGGCGGGGATTGCGACTTAT  2708

Query  421   GCCAATGCGTTGTCGAACCAGCTTGCCCCGCAAGAGGGGCTGGTTGCGGCGCAACACAGC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2707  GCCAATGCGTTGTCGAACCAGCTTGCCCCGCAAGAGGGGCTGGTTGCGGCGCAACACAGC  2648

Query  481   CTGACTTTCGCCGCCAATGGCTGGGTCGAACCAGCAACCGCGCCCAACTTTGGCCCGCTC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2647  CTGACTTTCGCCGCCAATGGCTGGGTCGAACCAGCAACCGCGCCCAACTTTGGCCCGCTC  2588

Query  541   AAGGTATTTTACCCCGGCCCCGGCCACACCAGTGACAATATCACCGTTGGGATCGACGGC  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2587  AAGGTATTTTACCCCGGCCCCGGCCACACCAGTGACAATATCACCGTTGGGATCGACGGC  2528

Query  601   ACCGACATCGCTTTTGGTGGCTGCCTGATCAAGGACAGCAAGGCCAAGTCGCTCGGCAAT  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct  2527  ACCGACATCGCTTTTGGTGGCTGCCTGATCAAGGACAGCAAGGCCAAGTCGCTCGGCAAT  2468

Query  661   CTCGGTGATGCCGACACTGAGCACTACGCCGCGTCAGCGCGCGCGTTTGGTGCGGCGTTC  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2467  CTCGGTGATGCCGACACTGAGCACTACGCCGCGTCAGCGCGCGCGTTTGGTGCGGCGTTC  2408

Query  721   CCCAAGGCCAGCATGATCGTGATGAGCCATTCCGCCCCCGATAGCCGCGCCGCAATCACT  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2407  CCCAAGGCCAGCATGATCGTGATGAGCCATTCCGCCCCCGATAGCCGCGCCGCAATCACT  2348

Query  781   CATACGGCCCGCATGGCCGACAAGCTGCGCTGA  813
             |||||||||||||||||||||||||||||||||
Sbjct  2347  CATACGGCCCGCATGGCCGACAAGCTGCGCTGA  2315
```

FIGURE 3
Gene mutation alignment.

Hypervirulence genes

The emergence of hypervirulent antimicrobial resistant bacteria has led to increased concern, as hypervirulence genes have been known to lead to more invasive and life-threatening illnesses. Hypervirulence and antimicrobial resistance were considered to be two divergent evolutionary pathways. However, in recent years organisms harboring both hypervirulence and antimicrobial resistance genes have emerged (38). Since DCLS added GAMMA to AMR analysis in February of 2022, of the 234 isolates analyzed the following hypervirulence genes were detected: 7 iroB-6, 3 iroB-23, 10 iucA-45, 1 iucA-18, 1 iucA-1, and 2 rmpA2–3. These genes were identified using a custom-database provided by CDC. These hypervirulence genes were all present in isolates that also tested positive for at least one carbapenem-resistance gene. Hypervirulence genes were originally described in *K. pneumoniae* isolates. However, the majority of the hypervirulence genes identified by DCLS since implementing GAMMA have been found in *E. cloacae* and *Escherichia coli*. One iroB-6 isolate was part of a multistate NDM + *E. cloacae* cluster found using the tiered method for surveillance.

Discussion

NGS has provided a higher-resolution method for identifying and tracking resistance and hypervirulence genes of concern, as well as performing a critical role in the epidemiological investigations of *Candida auris* and carbapenamase-producing organisms. These techniques allow epidemiologists to study epidemiological links between microorganisms. The lack of a centralized national database for CRO genomic epidemiology has stymied proactive surveillance-based detection of possible clusters of interest across multiple facilities, temporally disparate cases, and prolonged time frames. The methodology described herein harnesses a publicly available data repository that provides centralized and integrated bacterial pathogen genomic comparisons for cluster prediction. Notification tools available through NCBI can alert laboratorians and epidemiologists to matches to jurisdictional isolates, as they are identified in the Pathogen Detection algorithm. Further interrogating putative clusters with a within data-set reference can help to further discern the extent of genomic differences which serve as a proxy for likelihood of transmission of a common infectious bacterial strain. NGS has been used many times to assist Virginia Department of Health epidemiologists in the quest to stop the spread of disease and antimicrobial resistance.

In 2019, an outbreak of KPC *Pseudomonas aeruginosa* in Southwest Virginia at an acute care hospital was investigated. To determine if there was spread within the facility, screening was conducted at the hospital as well as an infection prevention and control assessment. Epidemiologists found a total of 2 cases within the facility and determined there was no spread outside of the facility. The investigation was considered closed.

In September of 2022, another case of KPC *Pseudomonas aeruginosa* was discovered in the same acute care hospital and was believed to be an isolated case. When the epidemiologists received the NGS surveillance results, they were able to determine that the case was 0 SNPs apart from the 2 cases in 2019. This information shifted the investigation and encouraged the team to investigate the possibility of sustained reservoirs within the hospital itself.

The source of the outbreak has yet to be determined; the investigation is still ongoing. NGS has enabled the team of epidemiologists to gain insight into the linkage between these three cases, whereas before it was thought that the cases were unrelated. NGS gives epidemiologists an extra tool to be able to stop multi-drug resistant organisms and protect some of our most vulnerable populations. By using NGS to elucidate linkages across outbreaks and identify the presence of resistance genes adds another line of defense to the arsenal of public health.

This example demonstrates the value of adding NGS surveillance to the DCLS microbiology workflow. These results can be used by epidemiologists to improve the prevention and control of these highly resistant infectious pathogens. In addition, surveillance enables the tracking and identification of novel and emerging resistance genes and pathogens within our region. Limitations to surveillance using NGS include the dependence on hospital compliance to CRO submission laws. Funding also limits sequencing all CRO isolates which may leave gaps in tracking the spread of AMR. In addition, not all isolates are submitted to NCBI, and metadata can be lacking. Deidentification can prevent linking patients from Virginia tested in other states due to cross border healthcare or reference laboratory testing. Multistate outbreaks can be difficult for follow up due to multiple health departments and public health laboratories involvement. Lack of standardization of AMR tools and databases also inhibits comparison of results from one source to another. Despite the limitations, implementing NGS surveillance over the last year has improved awareness and understanding of carbapenem-resistant organisms and their resistance genes within Virginia for both the public health laboratory and the health department. This raised awareness has shown the need for plasmid genomics to track the spread of plasmids and resistant genes between bacterial species and is the focus of future work at DCLS.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/2022EP-00101, https://www.ncbi.nlm.nih.gov/2022EP-00107.

Author contributions

KG and RS conceptualized the study, analysis plan, and analyzed the data. CD performed sequencing and wet lab data analysis. AO'R contributed to writing the draft and provided data. LT provided technical oversight and contributed to writing and editing the manuscript. All authors contributed to writing the final manuscript.

Funding

The work described herein was funded by the Centers for Disease Control and Prevention Epidemiology and Laboratory

Capacity for Prevention and Control of Emerging Infectious Diseases Cooperative Agreement with the National Center for Emerging and Zoonotic Infectious Diseases (5NU50CK000555-04-00).

Acknowledgments

Virginia's Division of Consolidated Laboratory Services Microbial Reference Group and Sequencing and Bioinformatics Group, and Virginia Department of Health Healthcare-Associated Infections (HAI) and Antimicrobial Resistance (AR) Group. Logan Fink for technical oversite and recommendations on bioinformatics processes. A special thanks to manuscript editors at both DCLS and Frontiers.

Conflict of interest

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

Publisher's note

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

References

1. Centers for Disease Control Prevention. *Antibiotic Resistance Threats in the United States 2019*. Atlanta, Georgia: US Department of Health and Human Services. (2019). Available online at: https://www.cdc.gov/drugresistance/pdf/threats-report/2019-ar-threats-report-508.pdf (accessed March 27, 2023).

2. The Federal Task Force on Combating Antibiotic-Resistant Bacteria. *U. S. Department of Health and Human Services*. (2020). Available online at: https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//196436/CARB-National-Action-Plan-2020-2025.pdf (accessed March 27, 2023).

3. Centers for Disease Control and Prevention. *About the AR Lab Network*. (2023). Available online at: https://www.cdc.gov/drugresistance/ar-lab-networks/domestic.html (accessed March 1, 2023).

4. StaPH-B Toolkit. (2019) [Source Code]. Available online at: https://github.com/StaPH-B/staphb_toolkit (accessed March 27, 2023).

5. StaPH-B. StaPH-B Consortium. (n.d.). Available online at: https://staphb.org/ (accessed March 6, 2023).

6. Centers for Disease Control Prevention. *CDC Antibiotic Resistance Laboratory Network: Whole Genome Sequencing of Healthcare-Associated Infection/Antibiotic Resistant Pathogens in State Local Public Health Laboratories*. Atlanta, Georgia: U. S. Department of Health and Human Services. (2020).

7. The NCBI Pathogen Detection Project [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. (2016). Available online at: https://www.ncbi.nlm.nih.gov/pathogens/ (accessed March 1, 2023).

8. Catho G, Martischang R, Boroli F, Chraïti MN, Martin Y, Koyluk Tomsuk Z, et al. Outbreak of *Pseudomonas aeruginosa* producing VIM carbapenemase in an intensive care unit and its termination by implementation of waterless patient care. *Crit Care*. (2021) 25:301. doi: 10.1186/s13054-021-03726-y

9. Zhou K, Lokate M, Deurenberg RH, Tepper M, Arends JP, Raangs EG, et al. Use of whole-genome sequencing to trace, control and characterize the regional expansion of extended-spectrum β-lactamase producing ST15 *Klebsiella pneumoniae*. *Sci Rep*. (2016) 6:20840. doi: 10.1038/srep20840

10. Pierce VE, Simner PJ, Lonsway DR, Roe-Carpenter DE, Johnson JK, Brasso WB, et al. Modified carbapenem inactivation method for phenotypic detection of carbapenemase production among enterobacteriaceae. *J Clin Microbiol*. (2017) 7:31. doi: 10.1128/JCM.00193-17

11. Center for Disease Control Prevention. *CDC AR Laboratory Network: Guidance for Testing CRE & CRPA in State Local Public Health Laboratories*. Atlanta, Georgia: U. S. Department of Health and Human Services (2021)

12. Qiagen. *QIAamp DNA Mini Blood Mini Handbook, 5th Edition*. (2016). Available online at: https://www.qiagen.com/us/resources/resourcedetail?id=62a200d6-faf4-469b-b50f-2b59cf738962&lang=en (accessed March 27, 2023).

13. Centers for Disease Control Prevention. *PulseNet Nextera DNA Flex Library Preparation*. Atlanta, Georgia: U. S. Department of Health and Human Services (2020).

14. Centers for Disease Control Prevention. *CDC Antibiotic Resistance Laboratory Network: Whole Genome Sequencing of Healthcare Associated Infection/Antibiotic Resistant Pathogens in State Local Public Health Laboratories*. Atlanta, Georgia: U. S. Department of Health and Human Services (2022).

15. Illumina. *Illumina DNA Prep Reference Guide*. (2020). Available online at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/illumina-dna-prep-reference-guide-1000000025416-10.pdf (accessed March 1, 2023).

16. Illumina. *MiSeq System Guide for Local Run Manage*r. (2019). Available online at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-for-local-run-manager-15027617-05.pdf (accessed March 27, 2023).

17. Illumina. *Does my sequencing run look good?* Available online at: https://support.illumina.com/bulletins/2019/10/does-my-sequencing-run-look-good-.html (accessed March 1, 2023).

18. Libuit KG. *StaPH-B Toolkit Tredegar [Source Code]*, (2019). Available online at: https://github.com/StaPH-B/staphb_toolkit (accessed February 27, 2022).

19. U.S. National Library of Medicine. (n.d.). *Genome list—genome—NCBI. National Center for Biotechnology Information*. Available online at: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/ (accessed March 3, 2023).

20. Seemann T. 2020. *Shovill: Assemble bacterial isolate genomes from Illumina paired-end reads. [Source Code]*. Available online at: https://github.com/tseemann/shovill (accessed March 6, 2023).

21. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. (2016) 17:132. doi: 10.1186/s13059-016-0997-x

22. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res*. (2023) 51:D29-D38. doi: 10.1093/nar/gkac1032

23. The NCBI Pathogen Detection Project. *Pathogen Detection Help Document. 2021*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available online at: https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/#data-processing (accessed March 27, 2023).

24. Feldgarden M, Brover V, Fedorov B, Haft DH, Prasad AB, Klimke W. Curation of the AMRFinderPlus databases: applications, functionality, and impact. *Microb Genom*. (2022) 8:mgen000832. doi: 10.1099/mgen.0.000832

25. St. Jacques RM. *StaPH-B Toolkit Hickory [Source Code]*. (2019). Available online at: https://github.com/StaPH-B/staphb_toolkit (accessed February 27, 2023).

26. Florek K, Shockey A. *StaPH-B Toolkit Dryad [Source Code]*. (2019). Available online at: https://github.com/StaPH-B/staphb_toolkit (accessed April 22, 2022).

27. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, et al. CFSAN SNP pipeline: An automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Computer Sci*. (2015) 1:15. doi: 10.7717/peerj-cs.20

28. Stanton RA, Vlachos N, Halpin AL. GAMMA: a tool for the rapid identification, classification and annotation of translated gene matches from sequencing data. *Bioinformatics*. (2022) 38:546–548. doi: 10.1093/bioinformatics/btab607

29. Pettengill JB, Markell A, Conrad A, Carleton HA, Beal J, Rand H, et al. A multinational listeriosis outbreak and the importance of sharing genomic data. *Lancet Microbe*. (2020) 1:e233–e234. doi: 10.1016/S2666-5247(20)30122-1

30. Centers for Disease Control and Prevention. *PulseNet*. (2021). Available online at: https://www.cdc.gov/pulsenet/index.html (accessed March 27, 2023).

31. U.S. National Library of Medicine. (n.d.). About the NCBI *pathogen detection system—pathogen detection—NCBI. National Center for Biotechnology Information.*

Available online at: https://www.ncbi.nlm.nih.gov/pathogens/about/ (accessed March 1, 2023).

32. U.S. National Library of Medicine. (n.d.). *Data Processing Pipeline.* Available online at: https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/#data-processing (accessed March 9, 2023).

33. Saltykova A, Mattheus W, Bertrand S, Roosens NHC, Marchal K, De Keersmaecker, et al. Detailed evaluation of data analysis tools for subtyping of whole genome sequencing: *neisseria meningitides* as a proof of concept. *Front Microbiol.* (2019) 10:2897. doi: 10.3389/fmicb.2019.02897

34. Lesho E, Clifford R, Onmus-Leone F, Appalla L, Snesrud E, Kwak Y, et al. The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. department of defense. *PLOS ONE.* (2016) 6:24. doi: 10.1371/journal.pone.0155770

35. Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuestra C, García-González N, Mejía L, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput Biol.* (2021) 17:78. doi: 10.1371/journal.pcbi.1008678

36. Hanussek M, Bartusch F, Krüger J. Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources. *PLOS Computational Biol.* (2021) 17:e1009244. doi: 10.1371/journal.pcbi.1009244

37. National Center for Biotechnology Information Pathogen Detection. *How to Request New Alleles for Beta-Lactamase, MCR, and Qnr Genes*. Available online at: https://www.ncbi.nlm.nih.gov/pathogens/submit-beta-lactamase/ (accessed December 1, 2022).

38. Lan P, Jiang Y, Zhou J, Yu Y. A global perspective on the convergence of hypervirulence and carbapenem resistance in *Klebsiella pneumoniae. J Glob Antimicrob Resist.* (2021) 25:26–34. doi: 10.1016/j.jgar.2021.02.020

frontiers | Frontiers in Public Health

TYPE Case Report
PUBLISHED 22 June 2023
DOI 10.3389/fpubh.2023.1197149

Check for updates

OPEN ACCESS

EDITED BY
Shangxin Yang,
University of California, Los Angeles,
United States

REVIEWED BY
Evann Hilt,
University of Minnesota Twin Cities,
United States
Huan Dong,
University of California, Los Angeles,
United States

*CORRESPONDENCE
Fang Liu
✉ Liufang_cd424@126.com

RECEIVED 30 March 2023
ACCEPTED 30 May 2023
PUBLISHED 22 June 2023

CITATION
Chang L, Che G, Yang Q, Lai S, Teng J, Duan J, Liu T and Liu F (2023) *Leishmania donovani* visceral leishmaniasis diagnosed by metagenomics next-generation sequencing in an infant with acute lymphoblastic leukemia: a case report. *Front. Public Health* 11:1197149. doi: 10.3389/fpubh.2023.1197149

Leishmania donovani visceral leishmaniasis diagnosed by metagenomics next-generation sequencing in an infant with acute lymphoblastic leukemia: a case report

Li Chang[1,2], Guanglu Che[1,2], Qiuxia Yang[1,2], Shuyu Lai[1,2], Jie Teng[1,2], Jiaxin Duan[1,2], Ting Liu[1,2] and Fang Liu[1,2]*

[1]Department of Laboratory Medicine, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China, [2]Key Laboratory of Birth Defects and Related Diseases of Women and Children, Sichuan University, Ministry of Education, Chengdu, Sichuan, China

Background: Visceral leishmaniasis (VL) is a neglected vector-borne tropical disease caused by *Leishmania donovani* (*L. donovani*) and *Leishmania infantum* (*L. infantum*). Due to the very small dimensions of the protozoa impounded within blood cells and reticuloendothelial structure, diagnosing VL remains challenging.

Case presentation: Herein, we reported a case of VL in a 17-month-old boy with acute lymphoblastic leukemia (ALL). The patient was admitted to West China Second University Hospital, Sichuan University, due to repeated fever after chemotherapy. After admission, chemotherapy-related bone marrow suppression and infection were suspected based on clinical symptoms and laboratory test results. However, there was no growth in the conventional peripheral blood culture, and the patient was unresponsive to routine antibiotics. Metagenomics next-generation sequencing (mNGS) of peripheral blood identified *196123 L. donovani* reads, followed by *Leishmania* spp amastigotes using cytomorphology examination of the bone marrow specimen. The patient was given pentavalent antimonials as parasite-resistant therapy for 10 days. After the initial treatment, *356 L. donovani* reads were still found in peripheral blood by mNGS. Subsequently, the anti-leishmanial drug amphotericin B was administrated as rescue therapy, and the patient was discharged after a clinical cure.

Conclusion: Our results indicated that leishmaniasis still exists in China. Unbiased mNGS provided a clinically actionable diagnosis of a specific infectious disease from an uncommon pathogen that eluded conventional testing.

KEYWORDS

visceral leishmaniasis, *Leishmania donovani*, metagenomic next-generation sequencing, acute lymphoblastic leukemia, rapid diagnosis, case report

Introduction

Visceral leishmaniasis (VL) is a vector-borne protozoan neglected tropical disease (NTDs) caused by *Leishmania donovani* complex (*L. donovani* and *L. infantum*) and *L. donovani* (1–3). It is caused by an infection of blood cells in the lymphoid organs, primarily the spleen, bone marrow, and liver, and is fatal in more than 95% of untreated cases (4).

In China, 3,169 cases of VL have been reported, with ∼140–509 cases diagnosed per year between 2002 and 2011. VL is considered endemic in over 50 counties across 6 provinces/autonomous regions in western China, including Xinjiang, Gansu, Sichuan, Shaanxi, Shanxi, and Inner Mongolia (5–8). According to these data, leishmaniasis is not extinct and could potentially cause a public health problem in China.

Acute lymphoblastic leukemia (ALL) is the most frequent type of pediatric cancer, with an incidence of 5.4 per 100,000 cases in patients aged <15 years old (9). In addition, cases of leishmaniasis found in patients formerly diagnosed with various cancers and treated with long-term anti-cancer chemotherapy have been previously reported, clearly suggesting an overlap between leishmaniasis transmission and malignant disease (10).

The diagnosis of VL is based on detecting *Leishmania* amastigote parasites in bone marrow or spleen biopsies. However, the very small dimensions of the protozoa impounded within blood cells and reticuloendothelial structure makes diagnosing leishmaniasis challenging (11). Recently developed metagenomics next-generation sequencing (mNGS) analyses forego the use of specific primers or probes. Instead, the entirety of the DNA and/or RNA (after reverse transcription to cDNA) is sequenced, thus providing a practical approach for diagnosing rare, novel, and atypical infectious etiologies (12). In the following description, we reported a VL case in an ALL infant after chemotherapy diagnosed by mNGS and parasitological microscopy.

Case presentation

A 17-month-old boy was admitted to West China Second University Hospital, Sichuan University, in March of 2022 for the insidious onset of fever, ecchymosis of skin, anhelation, and pancytopenia. On admission, blood routine examination results were as follows: white blood cell (WBC) 2.5 × 10^9/L, hemoglobin (Hb) 60 g/L, platelet (PLT) 45 × 10^9/L, and immature granulocytes found in peripheral blood smears. Subsequently, bone marrow morphology, immunophenotyping, cytogenetics, and molecular genetics were carried out. Based on the above testing, the boy was diagnosed with B-cell acute lymphoblastic leukemia (B-ALL, L2, ETV6-PEX5 fusion gene positive, KRAS A146V and KRAS A146T mutation, IKZF1–8 heterozygous deletion, 45, XY, der (7; 12) (q10; q10)(5)/46, XY (15). Then, according to ALL-low risk (ALL-LR) of the Chinese Children's Cancer Group study ALL 2020 (CCCG-ALL-2020), conventional and continuous therapy was administered. After the remission induction regimen of 4 weeks, complete remission (CR) was reached. Subsequently, the child was supposed to receive consolidation therapy according to CCCG-ALL-2020 with three courses after 4 weeks of achievement of CR.

On October 2022, which was also the interval between the first consolidation treatments, the patient was hospitalized again for a repeated fever of 13 days and coagulopathy after the second cycle of consolidation chemotherapy. Laboratory tests are shown in Table 1. Color Doppler ultrasonic examination showed swollen liver and spleen. The above results suggested that the infant had chemotherapy-related bone marrow suppression, infection, and coagulation dysfunction. Vancomycin, imipenem, and voriconazole were used for empirical antibiotic therapy. Fresh frozen plasma, fibrinogen, and prothrombin complex were used for improving coagulation function. Seven days after therapy, the patient still had a fever (38.4–39.4°C), and his liver and spleen were enlarged. CRP and PCT levels were 125.5 mg/L and 3.99 mg/L, respectively (Table 1). No microorganisms were detected by blood culture.

Then, mNGS was carried out in peripheral blood. After DNA was extracted from 200 µl of peripheral blood, the DNA library was built and sequenced on Nextseq 550 platform (Illumina, USA). All human host DNA was filtered out, and the valuable reads were aligned to Microbial Genome Databases (ftp://ftp.ncbi.nlm.nih.gov/genomes/) using BWA. Finally, a number of 196123 special reads of *L. donovani* were detected, and the coverage of the genome and relative abundance of *L. donovani* was 39.43 and 97.9%, respectively (Figure 1), which was indicative of *L. donovani* infection. In addition, 1 special read of *Klebsiella pneumoniae* was detected and defined as a background bacteria. Examination of patients' demographic information revealed the following: ever since birth, he resided in Jiuzhaigou county of Sichuan province, which is known as an endemic leishmaniasis region. Subsequently, microscopy of the bone marrow showed a larger number of *Leishmania* spp amastigotes, while phagocytic phenomena of histiocytes were found in all smears (Figure 2). All bone marrow smears from confirmed leukemia were reviewed, revealing no *Leishmania* spp amastigote in microscopy. Furthermore, RK39 antigens were also positive by immunochromatography. Thus, VL was diagnosed, and sodium stibogluconates were used as an anti-leishmanial drug. Nevertheless, after 10 days of anti-leishmanial therapy, 356 special reads of *L. donovani* were detected in peripheral blood by mNGS (Figure 3), and *Leishmania* spp amastigotes were also observed in the bone marrow. Subsequently,

TABLE 1 Laboratory test results of a patient during diagnosis and treatment.

Enrollment time	WBC (×10^9/L)	N (%)	L (%)	E (%)	Hb (g/L)	PLT (×10^9/L)	PT (S)	APTT (S)	Fg (mg/dL)	CRP (mg/L)
On admission	1.0	60.4	29.5	0	85	8	15.5	43.6	105	76.9
7 days after antibiotic treatment	12.2	91.0	3.0	0.1	122	39	12.7	34.2	200	125.5
10 days after sodium stibogluconate treatment	1.5	74.7	9.3	0	81	159	10.7	25.4	202	7.2
14 days after Amphotericin B	2.8	39.1	35.0	0.2	110	182	/	/	/	0.5

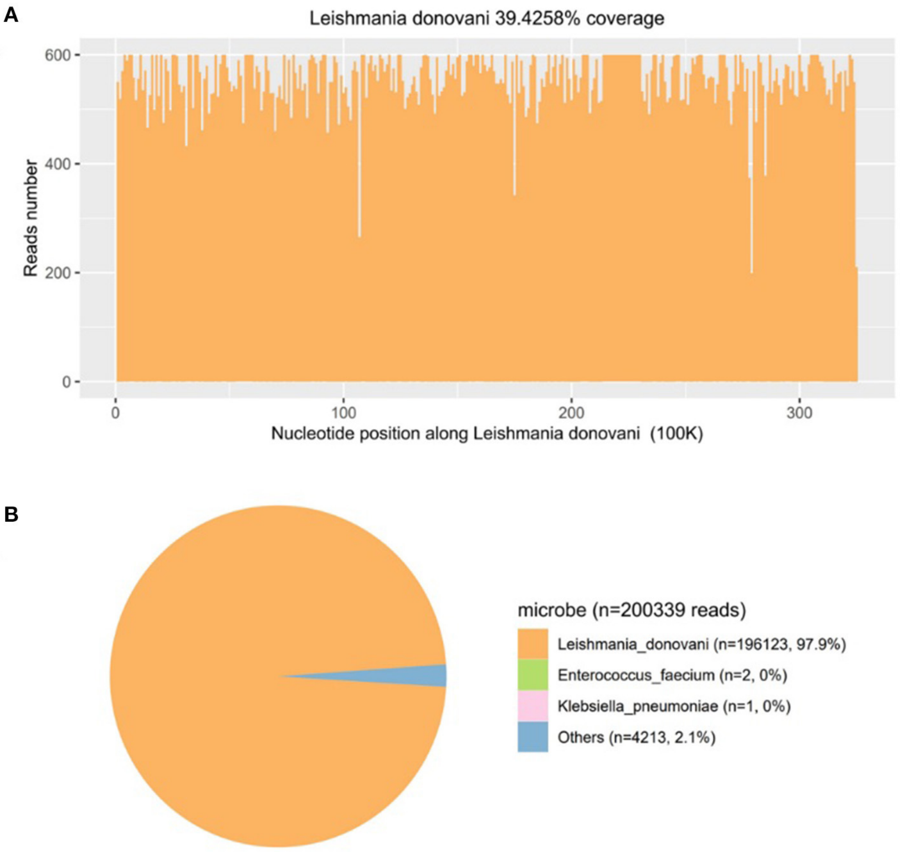


FIGURE 1
The diagnosis of *Leishmania* infection by metagenomics next-generation sequencing (mNGS). **(A)** Mapping of *Leishmania donovani* reads on the genome. **(B)** Distribution of pathogenic microorganisms reads in the absence of human, others, and unclassified reads.

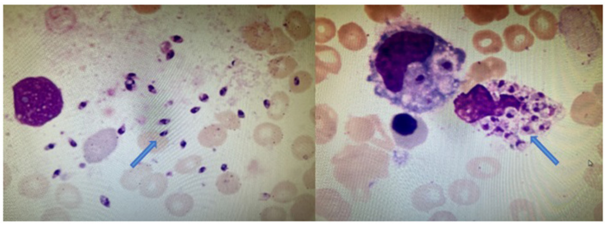


FIGURE 2
Bone marrow cytology of this patient. Arrowheads show the *Leishmania* spp amastigotes in extracellular and phagocyte, which are oval and 2.9–5.7 × 1.8–4.0 μm in size (Wright's stain, ×1,000).

the anti-leishmanial drug Amphotericin B was used as rescue therapy. After completion of therapy, there was no *Leishmania* spp amastigote in the bone marrow. Finally, the child was discharged 56 days after admission. He was subsequently referred to the hematological clinic for leukemia and follow-up. On February 2023, he was readmitted to the hospital for chemotherapy, when mNGS test for peripheral blood was performed, and no reads of *L. donovani* were detected.

Discussion

In the present study, we described a case of an infant diagnosed with ALL on admission to the hospital with repeated fever and coagulopathy after chemotherapy. A specific pathogen infection was suspected after collecting demographic information and learning about the history of lifelong residence in the forest region in Sichuan province, clinical symptoms, laboratory test results, and treatment history. Subsequently, the diagnosis of VL was definitely confirmed by mNGS. The patient was treated with pentavalent-Sb with adequate dosage and duration, and mNGS detected 356 *L. donovani* reads from the patient's plasma sample. Anti-leishmanial drug amphotericin B was subsequently administrated as rescue therapy. To the best of our knowledge, this is a rare report of leishmaniasis diagnosed by mNGS in leukemia, which provides a valuable reference for VL diagnosis and therapy follow-up.

Due to its wide geographic distribution, leishmaniosis constitutes a major public health problem. It is the second most prevalent pathogen among parasitic diseases. Hepatosplenomegaly, anemia, fever, cachexia, and leucopenia are all symptoms of this kind of VL, which can be significantly more dangerous (13). Factors that negatively impair the immune response, such as malnutrition or AIDS, are known to increase the risk of acquiring the infection

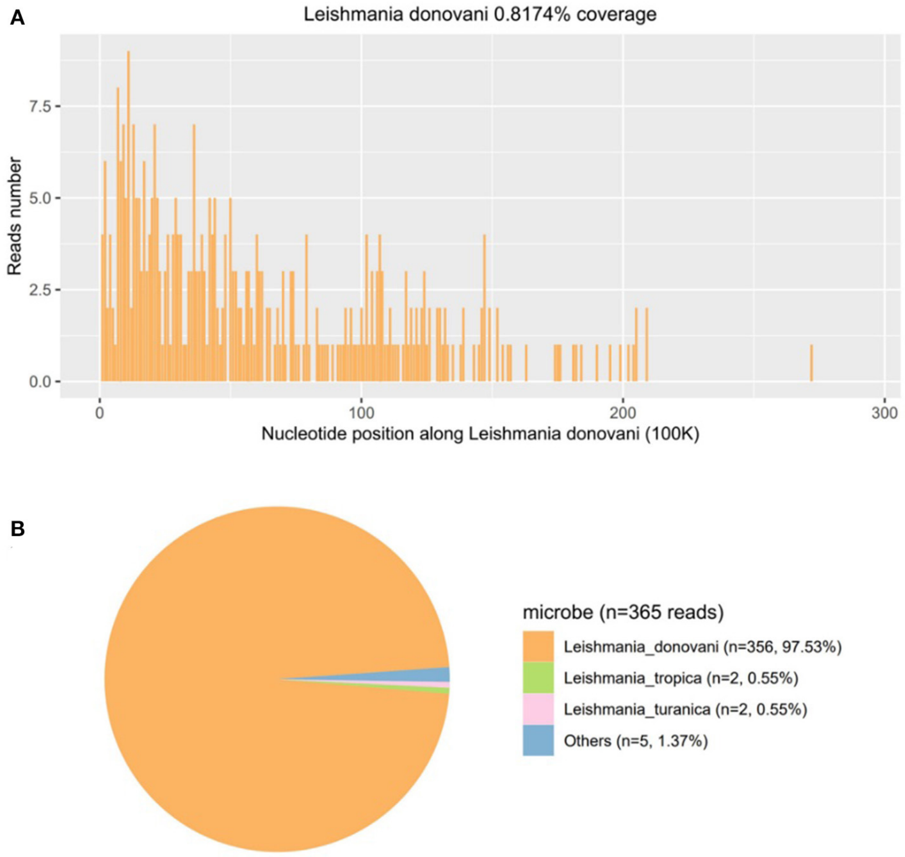


FIGURE 3
The follow-up diagnosis of *Leishmania* infection after using pentavalent antimonials 10 days by mNGS. **(A)** Mapping of *Leishmania donovani* reads on the genome. **(B)** Distribution of pathogenic microorganisms reads in the absence of human, others, and unclassified reads.

and result in more severe manifestations (14). Previous studies have reported that VL is a frequent opportunistic infection in HIV-infected immunodeficient individuals that is very rarely found in cancer patients (10, 13). Although immune suppression by treatments or diseases has been rarely described as a risk factor for VL, the most common underlying cause of immunodeficiency in patients with VL are hematological malignancies apart from HIV infection (15). Nonspecific manifestations such as fever, fatigue, hepatosplenomegaly, hepatosplenomegaly, and weight loss may be attributed to malignancy or related treatment, which is difficult to diagnose in patients with tumors (13). Considering the risk of infection, there is a semiquantitative interaction of 2 factors, i.e., epidemiological exposures and the net state of immunosuppression. In this study, the boy resided since birth in an endemic leishmaniasis region of Sichuan province, and anti-leukemia chemotherapy resulted in immunosuppression. Several studies reported a possible association between *Leishmania* infection and cancer (16). Although local immune suppression induced by malignant disorders may promote leishmaniasis development, it is more likely that immunosuppression induced by long-term anti-cancer chemotherapy is responsible for parasite expansion (16).

As *L. donovani* is a specific pathogen that is not commonly present in the environment, patient's epidemiological history and nonspecific manifestations may be easily overlooked. Also, serological or polymerase chain reaction (PCR) reagents for this pathogen are not routinely prepared in the laboratory, which may delay the diagnosis of this infection. Some VL cases have been misdiagnosed as autoimmune hepatitis, ALL, and malignant lymphoma. They can also be asymptomatic, occur in unusual locations, or be clinically or microbiologically refractory (10, 15, 17, 18). In the present case, the symptoms of fever, fatigue, and hepatosplenomegaly were attributed to the ALL or related treatment, and the absence of amastigotes on repeat bone marrow smears to leukemia diagnosis and surveillance, which is why the infection of *L. donovani* was initially ignored. Finally, the infection was definitely confirmed by mNGS as an unknown and refractory infection. The diagnosis of *Leishmania* infection is based on detecting *Leishmania* amastigote, and various diagnostic techniques were used in making the diagnosis. Most studies used combined immunological methods, while others used plot molecular and parasitological tests. Some cases are also challenging to diagnose due to the low parasite load and low levels of antibodies (19). As an unbiased approach to the detection of pathogens, mNGS has allowed crossing the divide from microbial research to diagnostic microbiology, overcoming limitations of current diagnostic tests and allowing for hypothesis-free, culture-independent pathogen detection directly from clinical specimens (20, 21). In addition, for infectious diseases, collecting patient medical history, especially epidemiological history, is very important for diagnosis, as it can help us quickly adopt appropriate detection methods to identify the pathogen. However, in this case, the history of leukemia has its specificity, which has led to the neglect of typical bone marrow microscopy and medical history collection. Nonetheless, future clinical studies are needed to further confirm the value of mNGS in diagnosing *Leishmania* infection.

In summary, VL should be considered a potential opportunistic infection in patients with hematologic malignancies, especially in immunosuppressed patients living in or having visited areas where the disease is endemic. Unbiased mNGS may provide a clinically actionable diagnosis of a specific infectious disease from an uncommon pathogen, eluding conventional testing for weeks after the initial presentation.

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.

Ethics statement

The studies involving human participants were reviewed and approved by Medicine Ethics Committee of West China Second University Hospital, Sichuan University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

Author contributions

Conception and design: FL. Provision of study materials or patients: LC. Collection and assembly of data: LC, GC, and QY. Data analysis and interpretation: SL, JT, and JD. Manuscript writing and final approval of manuscript: all authors. All authors contributed to the article and approved the submitted version.

Funding

This research received funding support by Key R & D Projects of Science and Technology Department of Sichuan Province, 2020YFS0106 and 2023YFS0186.

Conflict of interest

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

Publisher's note

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

References

1. Burza S, Croft SL, Boelaert M. Leishmaniasis. *Lancet.* (2018) 15:951–70. doi: 10.1016/S0140-6736(18)31204-2

2. Alvar J, Yactayo S, Bern C. Leishmaniasis and poverty. *Trends Parasitol.* (2006) 22:552–7. doi: 10.1016/j.pt.2006.09.004

3. Leta S, Dao THT, Mesele F, Alemayehu G. Visceral leishmaniasis in Ethiopia: an evolving disease. *PLoS Negl Trop Dis.* (2014) 4:8. doi: 10.1371/journal.pntd.0003131

4. Wilhelm TJ. Visceral Leishmaniasis. *Chrirurg.* (2019) 90:833–7. doi: 10.1007/s00104-019-0994-1

5. Lun ZR, Wu MS, Chen YF, Wang JY, Zhou XN, Liao LF, et al. Visceral leishmaniasis in china: an endemic disease under control. *Clin Microbiol Rev.* (2015) 28:987–1004. doi: 10.1128/CMR.00080-14

6. Etebari K, Hegde S, Saldaña MA, Widen SG, Wood TG, Asgari S, et al. Global transcriptome analysis of *Aedes aegypti* mosquitoes in response to Zika Virus infection. *mSphere.* (2017) 22:2. doi: 10.1128/mSphere.00456-17

7. Wang JY, Feng Y, Gao CH, Jin CF, Chen SB, Zhang CJ, et al. Asymptomatic *Leishmania* infection in human population of Wenxian County, Gansu Province. *Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi.* (2007) 25:62–4.

8. Guan LR, Zuo XP, Yimamu. Reemergence of visceral leishmaniasis in Kashi Prefecture. *Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi.* (2003) 21:285.

9. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat Rev Cancer.* (2018) 18:471–84. doi: 10.1038/s41568-018-0015-6

10. Kopterides P, Mourtzoukou EG, Skopelitis E, Tsavaris N, Falagas ME. Aspects of the association between leishmaniasis and malignant disorders. *Trans R Soc Trop Med Hyg.* (2007) 101:1181–9. doi: 10.1016/j.trstmh.2007.08.003

11. Sreedharan V, Rao KVB. Protease inhibitors as a potential agent against visceral leishmaniasis: a review to inspire future study. *Braz J Infect Dis.* (2023) 27:1. doi: 10.1016/j.bjid.2022.102739

12. Charles Y, Chiu SA. Miller clinical metagenomics. *Nat Rev Genet.* (2019) 20:341–55. doi: 10.1038/s41576-019-0113-7

13. Albrecht H, Sobottka I, Emminger C, Jablonowski H, Just G. Stoehr A, et al.Visceral leishmaniasis emerging as an important opportunistic infection in HIV-infected persons living in areas nonendemic for *Leishmania donovani*. *Arch Pathol Lab Med.* (1996) 120:189–98.

14. Desjeux P. Leishmaniasis: current situation and new perspectives. *Comp Immunol Microbiol Infect Dis.* (2004) 27:305–18. doi: 10.1016/j.cimid.2004.03.004

15. Fernández-Guerrero ML, Aguado JM, Buzón L, Barros C, Montalbán C, Martín T, et al. Visceral leishmaniasis in immunocompromised hosts. *Am J Med.* (1987) 83:1098–102. doi: 10.1016/0002-9343(87)90948-X

16. Fishman JA. Infections in immunocompromised hosts and organ transplant recipients: essentials. *Liver Transpl.* (2011) 17:3. doi: 10.1002/lt.22378

17. Dalgic B, Dursun I, Akyol G. A case of visceral leishmaniasis misdiagnosed as autoimmune hepatitis. *Turk J Gastroenterol.* (2005) 16:52–3.

18. Jones SG, Forman KM, Clark D, Myers B. Visceral leishmaniasis misdiagnosed as probable acute lymphoblastic leukaemia. *Hosp Med.* (2003) 64:308–9. doi: 10.12968/hosp.2003.64.5.1767

19. Neitzke-Abreu HC, Venazzi MS, Bernal MV, Reinhold-Castro KR, Vagetti F, Mota CA, et al. Detection of DNA from *Leishmania* (Viannia): accuracy of polymerase chain reaction for the diagnosis of cutaneous leishmaniasis. *PLoS ONE.* (2013) 5:e0062473. doi: 10.1371/journal.pone.0062473

20. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology, et al. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch Pathol Lab Med.* (2017) 141:776–86. doi: 10.5858/arpa.2016-0539-RA

21. Simner PJ, Miller S, Carroll KC. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. *Clin Infect Dis.* (2018) 10:778–88. doi: 10.1093/cid/cix881

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 29 June 2023
DOI 10.3389/fpubh.2023.1206056

OPEN ACCESS

EDITED BY
Peera Hemarajata,
Association of Public Health Laboratories,
United States

REVIEWED BY
Jennifer K. Spinler,
Baylor College of Medicine, United States
Hannah Gray,
University of California,
Los Angeles, United States

*CORRESPONDENCE
Kimberlee A. Musser
✉ kimberlee.musser@health.ny.gov

RECEIVED 14 April 2023
ACCEPTED 07 June 2023
PUBLISHED 29 June 2023

CITATION
Murphy SG, Smith C, Lapierre P, Shea J, Patel K, Halse TA, Dickinson M, Escuyer V, Rowlinson MC and Musser KA (2023) Direct detection of drug-resistant *Mycobacterium tuberculosis* using targeted next generation sequencing.
Front. Public Health 11:1206056.
doi: 10.3389/fpubh.2023.1206056

Direct detection of drug-resistant *Mycobacterium tuberculosis* using targeted next generation sequencing

Shannon G. Murphy, Carol Smith, Pascal Lapierre, Joseph Shea, Kruthikaben Patel, Tanya A. Halse, Michelle Dickinson, Vincent Escuyer, Marie Claire Rowlinson and Kimberlee A. Musser*

Wadsworth Center, New York State Department of Health, Albany, NY, United States

Mycobacterium tuberculosis complex (MTBC) infections are treated with combinations of antibiotics; however, these regimens are not as efficacious against multidrug and extensively drug resistant MTBC. Phenotypic (growth-based) drug susceptibility testing on slow growing bacteria like MTBC requires many weeks to months to complete, whereas sequencing-based approaches can predict drug resistance (DR) with reduced turnaround time. We sought to develop a multiplexed, targeted next generation sequencing (tNGS) assay that can predict DR and can be performed directly on clinical respiratory specimens. A multiplex PCR was designed to amplify a group of thirteen full-length genes and promoter regions with mutations known to be involved in resistance to first- and second-line MTBC drugs. Long-read amplicon libraries were sequenced with Oxford Nanopore Technologies platforms and high-confidence resistance mutations were identified in real-time using an in-house developed bioinformatics pipeline. Sensitivity, specificity, reproducibility, and accuracy of the tNGS assay was assessed as part of a clinical validation study. In total, tNGS was performed on 72 primary specimens and 55 MTBC-positive cultures and results were compared to clinical whole genome sequencing (WGS) performed on paired patient cultures. Complete or partial susceptibility profiles were generated from 82% of smear positive primary specimens and the resistance mutations identified by tNGS were 100% concordant with WGS. In addition to performing tNGS on primary clinical samples, this assay can be used to sequence MTBC cultures mixed with other mycobacterial species that would not yield WGS results. The assay can be effectively implemented in a clinical/diagnostic laboratory with a two to three day turnaround time and, even if batched weekly, tNGS results are available on average 15 days earlier than culture-derived WGS results. This study demonstrates that tNGS can reliably predict MTBC drug resistance directly from clinical specimens or cultures and provide critical information in a timely manner for the appropriate treatment of patients with DR tuberculosis.

KEYWORDS

mycobacterium, tuberculosis, drug susceptibility, resistance, targeted sequencing, nanopore

1. Introduction

Tuberculosis (TB) continues to be a major contributor to global infectious disease deaths, with an estimated 10.6 million cases and 1.6 million deaths worldwide in 2021 (1). TB patients are treated with combination drug regimens; however, the emergence of increasingly drug-resistant forms of *Mycobacterium tuberculosis* complex (MTBC) in recent decades necessitates the use of alternative therapies (2). Currently, strategies for therapy are still mostly decided based on culture-based phenotypic drug susceptibility testing (DST); however, MTBC DST requires weeks or months to complete due to the slow growth rate of this organism (3–7). During this time, patients can be prescribed ineffective drug regimens, leading to treatment failure or the promotion of drug resistance (DR) (8). The potential for these negative patient outcomes underscores the need for quicker methods to detect DR TB (DR-TB).

Molecular and sequencing-based assays offer a faster alternative for profiling DR in slow-growing organisms like *M. tuberculosis*. Commercially available molecular methods include Xpert MTB/RIF (Cepheid, Sunnyvale, CA) (9) and the GenoType MTBDRplus line probe assay (Hain Lifescience Nehren, Germany) (10), which detect mutations within specific "hot spot" regions to predict DR. These rapid diagnostics, endorsed by World Health Organization (WHO), have contributed to improved global detection of DR, particularly for the first-line drug rifampin (11–14). These assays, however, may miss mutations outside of the targeted "hot spot" regions and incur false negative results (15, 16), or, in rare circumstances, silent or neutral mutations may incur false positive results for DR (17, 18).

Sequencing-based methods provide greater resolution of these loci. Assays developed for the detection of DR include pyrosequencing (19–22) and Sanger sequencing (23) of individual targets; however, these methods are typically limited to single-plex reactions analyzing limited sections of DR determining loci. NGS assays offer more comprehensive DR profiles by identifying novel and high confidence DR-associated mutations throughout the genome (24, 25). Whole genome sequencing (WGS) assays, such as the one implemented by Wadsworth Center, identify high-confidence mutations that allow accurate prediction of phenotypic DR (26). These assays provide comprehensive DR profiling and the bioinformatic analysis can be routinely updated to include new loci and mutations in accordance with national and global WHO databases (27). WGS results can be generated before phenotypic DST is available (7, 26, 28, 29); however, most clinically validated WGS assays are performed on MTBC-positive cultures that can require several weeks of incubation (30).

Targeted NGS (tNGS) assays can further reduce the time required for comprehensive DR profiling by amplifying numerous loci directly from clinical specimens. Several tNGS assays for DR profiling have been described in the literature, including laboratory-developed assays (31–35) and the commercially available Genoscreen Deeplex (36, 37) and Ion AmpliSeq (38). These assays vary in a number of ways including the selection and size of targets, how multiplexed the PCR reactions are, and the sequencing platforms employed, which include Illumina (31, 35, 36), Ion Torrent (32, 34, 38), and Oxford Nanopore Technologies (33, 39, 40).

In this paper, we describe the design, validation, and implementation of a tNGS assay for direct DR profiling on MTBC-positive clinical specimens at the Wadsworth Center. This assay includes a simplified set up with two multiplexed PCR amplification reactions that target thirteen full-length loci implicated in DR to first- and second-line MTBC antimicrobials. The assay was optimized for sequencing on an Oxford Nanopore Technologies platform, enabling real-time analysis, a two-to-three-day turnaround time with typically <2 h of sequencing time, and a cost of less than $80 per sample. In addition to performing tNGS on primary specimens, this assay was found to be accurate and generated susceptibility profiles comparable to those currently obtained with our existing WGS assay, which can only be performed on cultured isolates. These results demonstrate that tNGS-based assays can provide a reliable and cost-effective tool for early detection of DR-TB and should be considered for implementation in public health and clinical laboratories with MTBC testing needs and resources.

2. Materials and methods

2.1. Sample preparation

Samples submitted for mycobacterial testing were handled in a BSL-3 laboratory. Sterile tissue specimens (e.g., lung tissue, lymph node tissue) were homogenized in sterile saline. Respiratory specimens (e.g., sputum, bronchial washes, and bronchoalveolar lavages) underwent digestion and decontamination to optimize mycobacteria recovery. This procedure uses a 3.5% sodium hydroxide solution to dissolve mucus, lyse organic material, and inactivate other bacteria. Following incubation, the solution was neutralized, bacteria were concentrated by centrifugation, and the pellet was resuspended in a buffer. Processed samples were used to inoculate liquid cultures (MGIT 960, BACTEC) and underwent differential staining and smear microscopy. Aliquots for molecular testing were heat inactivated (80°C for 1 h) before handling in a BSL-2 laboratory.

2.2. Direct smear microscopy for acid-fast bacilli

Processed primary specimens were stained using the Ziehl-Neelsen Carbol Fuchsin method according to manufacturer instructions (Remel Inc., San Diego, CA) and examined under a microscope for the presence of Acid-Fast Bacilli (AFB). Samples positive for AFB were further categorized based on the number of AFB observed, with numerous defined as >9 AFB per high power field (HPF-1000X) (++++), moderate as 1–9 AFB per HPF (+++), few as 1–9 AFB per 10 HPF (++), and rare as 1–9 AFB per 300 HPF (+). Smear negative samples are defined as those with no AFB observed.

2.3. Real-time PCR for MTBC detection

DNA was extracted via mechanical lysis with FastPrep24 (MP Biomedicals, Solon, Ohio) and tested for *M. tuberculosis* complex (MTBC) DNA using previously described real-time PCR assay (41). This multiplexed assay includes a single-copy (ext-RD9) and multi-copy (IS*6110*) target for MTBC detection and a target for *Mycobacterium avium* complex detection (ITS). All specimens included in this study were positive for MTBC DNA via a real-time PCR.

2.4. Whole genome sequencing

Samples included in this study were analyzed using a NYS-validated WGS assay as previously described (26). Briefly, a manual DNA extraction utilizing InstaGene reagent (Bio-Rad Laboratories, Hercules, CA), mechanical lysis, and centrifugation was performed on heat-killed isolates identified as MTBC-positive. Concentration of DNA was assessed using Qubit DNA fluorometry (Thermo Fisher Scientific, Waltham, MA) and samples were prepared for Illumina sequencing on a MiSeq or NextSeq instrument (Illumina, San Diego, CA). Results were analyzed using a clinically validated in-house developed bioinformatics pipeline that identifies high-confidence and unknown/novel mutations (26).

2.5. DNA extraction and controls for tNGS

For tNGS, an automated lysis and purification-based DNA extraction method (EZ1 Virus DSP Kit, Qiagen) was used to minimize DNA shearing. On this platform, 100 μL of specimen was extracted and eluted in 60 μL. Each run included a positive and negative control, consisting of 100 μL of *Mycobacterium bovis* BCG MGIT positive culture and 100 μL of sterile molecular grade water, respectively. Both extraction controls were processed in parallel with clinicals specimens and serve as reagent and sequencing controls for the entire tNGS assay.

2.6. Primer design and PCR

Thirteen primer sets were designed to amplify full-length genes (*rpoB*, *katG*, *mabA*, *inhA*, *embB*, *gyrA*, *gyrB*, *ethA*, *rrs*, *rpsL*, and *pncA*) and/or promoter regions (*oxyR-ahpC*, *mabA-inhA*, *embC-A*, *pncA*, and *eis*) implicated in DR to first- and second-line MTBC antimicrobials, including rifampin, isoniazid, ethambutol, pyrazinamide, fluoroquinolones, ethionamide, streptomycin, and kanamycin/amikacin (Supplementary Table S1). Possible primer pairs were generated using Primer3 (42) and checked for *in silico* interactions with ThermoFisher's Multiple Primer Analyzer Tool (ThermoFisher Scientific, Waltham, MA) (43). Primer sets were multiplexed into two PCR reactions referred to as "Pool A" and "Pool B" and primer concentrations were optimized to obtain balanced amplification of each target (Supplementary Table S2). Each 40 μL PCR reaction contained Long Amp Hot Start Taq Mastermix (New England Biolabs, Ipswich, MA), DMSO (5% final concentration), and 5 μL of template. PCR was run for 40 cycles (with a 3 min and 30 s extension time) according to manufacturer instructions. Amplicons were visualized via gel electrophoresis alongside a 1 kilobase ladder.

2.7. Library preparation for nanopore sequencing

PCR reactions for each sample were combined and prepared for sequencing using ligation-based reagents from Oxford Nanopore Technologies (ONT; Oxford, United Kingdom) and adapted protocols (44). An overview of library preparation steps is illustrated in Figure 1A. Briefly, amplicons were purified using AMPure XP (Beckman Coulter, Brea, CA) and quantified using a Qubit™ Flex Fluorometer (ThermoFisher Scientific, Waltham, MA). Samples were normalized for concentration prior to a two-step "spike-in" method for DNA end repair and barcode ligation (44). Barcoded products were purified using AMPure XP, followed by adapter ligation and a final AMPure XP clean-up. Final eluate concentrations were measured, samples were pooled in equal ratios, and the final library was diluted to a concentration of 35 ng/μL. A 12 μL volume of the library was loaded onto an R9 flow cell according to manufacturer instructions. The run was sequenced on either a MinION Mk1C or a GridION platform with high-accuracy base calling until approximately 50 k reads per sample were obtained. Flow cells were washed according to manufacturer instructions and reused only if the flow cell retained sufficient active pores (>450) and only with uniquely barcoded samples to limit potential cross-contamination.

2.8. Bioinformatic analysis

Oxford Nanopore Technologies sequencing data is analyzed in real-time using a custom bioinformatics pipeline (Figure 1B), akin to the NYS-validated WGS pipeline described in Shea et al. (26). The pipeline reads in each of the raw fastq files as they are generated using the MinKNOW interface on the instrument. Fastq files are demultiplexed using Guppy on a separate server via command line interface (CLI) with default parameters. Reads are combined into a final fastq file for each analyzed sample. The pipeline then assesses the taxonomic content of each file using Kraken2 (version 2.1.2) and the k2_standard_08gb_20220607 database (45). All non-*Mycobacterium* genus reads are filtered out for the rest of the downstream analyses. Reads are mapped to *Mycobacterium tuberculosis* H37Rv reference genome with minimap2 (version 2.24-r1122) (46) and amplicon primers sequences are hard-clipped from both ends using SAMtools (v 1.15.1) with ampliconclip (47). Finally, a high-quality consensus sequence is generated for each sample using SAMtools mpileup (47) with minimum mapping quality and base quality of Phred 30 and 12, respectively, and minimum depth of 10× and 60% allele agreement. Indels require 40× minimum depth and 55% allele agreement. In cases where indels are directly adjacent or inside homopolymeric regions of three or more identical bases, percent allele agreement is raised to 75%. If a position (variant or invariant) does not reach these requirements, it is assigned as 'N' on the consensus sequence. The pipeline identifies 86 high-confidence resistance mutations across the 51 positions listed in Supplementary Table S1, and notes novel/unknown mutations. The different cutoffs for single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) were empirically determined by assessing the different allele frequencies (AF) over several runs and determining the best AF cutoff to avoid calling any false positive SNP or INDEL variants.

3. Results

3.1. Validation of tNGS for clinical use

To validate the tNGS assay for clinical use, we assessed sensitivity, reproducibility, specificity, and accuracy. To assess sensitivity of tNGS on respiratory specimens, a culture of the *M. tuberculosis* reference strain H37Rv (ATCC 25618) was serially diluted and spiked into

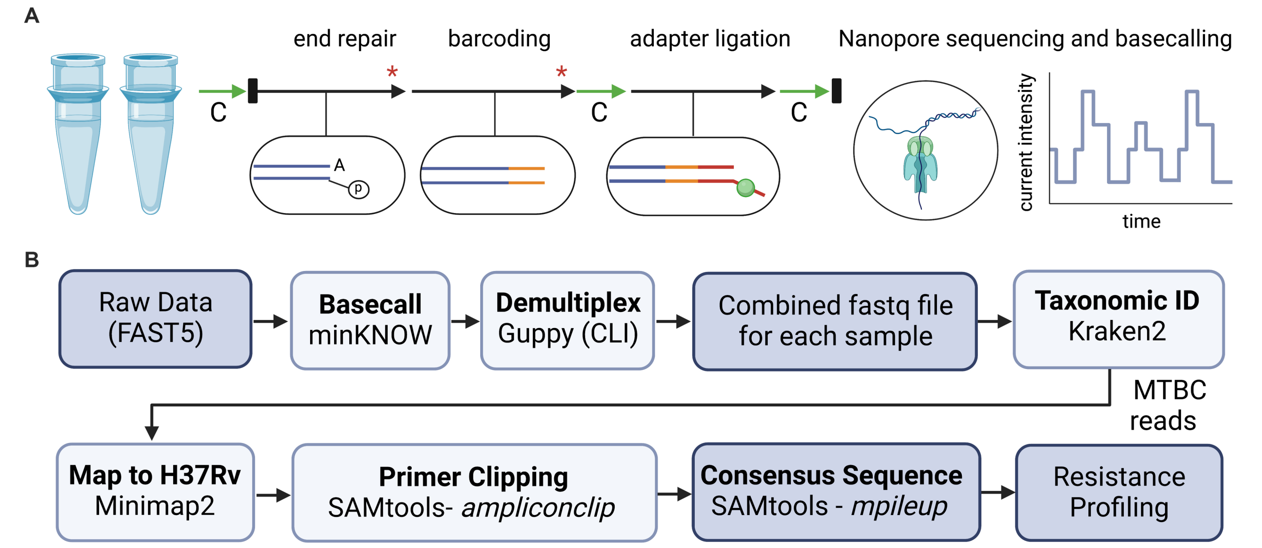


FIGURE 1
Overview of library preparation steps and bioinformatic analyses for tNGS nanopore sequencing. **(A)** For library preparation, two multiplex PCR reactions for each sample were combined and processed with AMPure bead-based clean-up steps (green arrows, "C"), enzymatic reactions (black arrows), dsDNA quantification via Qubit and normalization (black rectangles), and heat inactivation steps (red asterisk). **(B)** Bioinformatic tools used to analyze sequencing data and identify high confidence resistance mutations in MTBC. Diagrams created with BioRender.com.

processed negative sputa to determine the limit of detection (LOD). Average Ct-values for MTBC detection ranged from 24.2 to undetected and the concentration of *M. tuberculosis* in each PCR reaction ranged from 108 CFU to 0.00108 CFU (21,600 CFU to 0.216 CFU per mL). tNGS was performed on three replicate dilution series and sequenced to a total of approximately 80 k reads per sample. Quality control (QC) metrics were met for all targets (and corresponding drug classes) down to a lower detection limit of 0.108 CFU per reaction (Supplementary Table S3).

To measure reproducibility, three replicates of three smear positive specimens were processed in parallel (intra-assay) or on separate days (inter-assay). Results were concordant within and between runs, as shown in Supplementary Table S4. Specificity was tested against a panel of five organisms – including two mycobacteria (*Mycobacterium fortuitum* and *Mycobacterium abscessus*) and three other organisms common in sputa (*Klebsiella pneumoniae*, *Streptococcus pneumoniae*, and *Haemophilus influenzae*). No cross-reactivity was detected in this panel of organisms (Supplementary Table S5).

3.2. tNGS detection of drug resistance directly on respiratory specimens

To measure assay accuracy, tNGS was performed on a panel of 72 extracted primary specimens that were selected for their diverse mutations and drug resistance profiles. All specimens included in the panel were confirmed positive for MTBC DNA via real-time PCR. The panel consisted of 35 retrospective blinded samples and 37 prospective samples received over a period of 8 months (May 2022 to January 2023). The panel included predominantly sputa ($n = 58$, 81%) along with other respiratory specimens (i.e., bronchoalveolar lavages and bronchial washes) and rarer specimen types (i.e., lymph nodes and lung tissue). Specimens covered a range of MTBC concentrations (assessed by AFB smear and real-time PCR); most specimens included were AFB positive ($n = 65$, 90%), but five AFB negative samples and two untested samples were also included in the study (Supplementary Table S6).

The two multiplex PCR reactions were performed on the panel of specimens and amplification was confirmed by gel electrophoresis. Amplicons that could be visualized with ethidium bromide gel staining following PCR were present in 78% of the samples tested. High confidence resistance and unknown/novel mutations and DR profiles identified by tNGS were compared to those obtained from the NYS-validated WGS assay on isolates from the matched specimens. Profiles were defined according to CDC definitions: multidrug resistant (MDR; INH and RIF resistant), pre-extensively drug resistant (pre-XDR; INH, RIF, FQ), extensively drug resistant (XDR; INH, RIF, FQ, KAN/AMI). Resistance to other MTBC antimicrobials not meeting the criteria above is defined here as other mono- or poly-resistant (R). The results for each specimen are shown in Supplementary Table S6 and an aggregate summary is provided in Table 1. Of the MTBC-positive samples sequenced, tNGS correctly identified 44 pan-susceptible, 5 mono/poly-resistant, and 5 MDR, and 1 pre-XDR, and 1 XDR strain, all determined to have a DR profile identical to the WGS DR profile obtained from the culture isolate from the same case. At the mutation level, two tNGS reports identified additional unknown mutations in primary specimens that were not identified by WGS performed on cultured isolates. This raises the potential for tNGS to detect subpopulations in the primary clinical specimens. Overall, these results demonstrate that tNGS can accurately detect susceptible and DR forms of MTBC directly from primary specimens.

3.3. Primary specimen tNGS data quality

To evaluate data quality, samples were categorized based on the number of targets that met quality control thresholds defined by the analysis pipeline, either as complete susceptibility profiles (all 13 targets pass QC), partial susceptibility profiles (≥10 targets pass QC),

TABLE 1 Comparison of DR profiles identified by tNGS performed on primary specimens to WGS performed on matched MTBC-positive cultures.

		WGS (culture)				
		S	R	MDR	Pre-XDR	XDR
tNGS (Primary)	S	44	0	0	0	0
	R	0	5	0	0	0
	MDR	0	0	5	0	0
	Pre-XDR	0	0	0	1	0
	XDR	0	0	0	0	1
	Not sequenced	13	1	2	0	0
	Total	57	6	7	1	1

Profiles are categorized as pan-susceptible (S), mono- or poly-resistant (R), multidrug resistant (MDR), pre-extensively drug resistant (pre-XDR), and extensively drug resistant (XDR).

or no profile (not sequenced). In the panel of 72 primary specimens, tNGS produced 68% complete profiles, 10% partial profiles, and the remaining 22% were not sequenced due to PCR failure. These results indicate that targeted sequencing data can be obtained direct from primary specimens, although there is some variability in data quality.

To determine the factors that influence tNGS target failure and establish quality criteria for testing, the bacterial load in samples was estimated using AFB smear microscopy and MTBC real-time PCR Ct-values. Complete or partial profiles were obtained for 83% of the smear positive specimens tested ($n = 65$) (Figure 2A). Within the subset of smear positive samples, these percentages correlated with AFB smear results: 100% of samples with numerous AFB produced complete profiles, whereas complete or partial profiles were obtained for 93% of AFB moderate, 86% of AFB few, and 55% of AFB rare samples (Figure 2B). Of the five smear negative samples tested, only one specimen yielded a susceptibility profile; however, this sample had a low Ct-value uncharacteristic of a smear negative result (Supplementary Table S6). A smear could not be performed on several specimens due to specimen viscosity or insufficient volume for testing. These results indicate that target amplification is dependent on the quantity of AFB cells in the specimen and suggest that AFB smear-positive specimens are the most likely to yield complete susceptibility profiles.

Samples were also stratified by Ct-values derived from MTBC real-time PCR testing. Ct-values for the single-copy MTBC target (ext-RD9) ranged from 22.1 to 37.4 (or undetected) (Supplementary Table S6). Lower Ct-values yielded more complete tNGS sequencing results (Figure 2C); for values of 34.9 and below, 89% of samples yielded either complete or partial susceptibility profiles. In contrast, samples with Ct-values ≥35 were more prone to PCR failure (82%) and only two samples above this threshold produced a partial profile. Examination of Ct-values for IS*6110*, which is a multi-copy target and considered a more sensitive marker for MTBC detection, showed similar trends but with different ranges (Figure 2D). IS*6110* Ct-values ranged from 18.4 to 38.0 in the primary specimens tested (Supplementary Table S6). Samples with Ct-values ≤31.9 yielded either complete or partial susceptibility profiles (91%), whereas Ct-values ≥32 more were more prone to PCR failure (73%). These results indicate the quality of tNGS data is dependent on the amount of MTBC DNA present in the specimen and further suggests that quantification via real-time PCR may be used as a reliable metric for assessing sample quality for tNGS.

3.4. tNGS improves turnaround times

The ability of tNGS to generate comprehensive susceptibility profiles directly from a patient specimen has the potential to reduce turnaround times. A subset of 16 primary specimens with matched WGS results were used to calculate turnaround times; samples included in the analysis had tNGS performed as part of the routine testing algorithm (i.e., initiated within 1 week of MTBC detection) and yielded a positive MGIT culture suitable for WGS (Figure 3A). The average number of days required for MTBC detection via real-time PCR, tNGS (from extraction to result), MTBC isolation, and WGS (from extraction to result) are shown in Figure 3B.

Both tNGS and WGS samples were batched and run weekly. On average, tNGS results were available 10 days from sample receipt (or 7 business days if excluding weekends) (Figure 3B). This represents a 15 day reduction in turnaround time for tNGS versus WGS, with the improvement among samples ranging from 6 to 31 days (Figure 3C). Notably, one specimen included in this study was identified as XDR using tNGS and these results were available within 5 days from sample receipt, whereas culture-based WGS results were not available for an additional 3 weeks. These differences in turnaround time can largely be attributed to the incubation period required to obtain an AFB-positive culture and subsequent characterization (average 12.2 days); however, we also found that processing times – from sample extraction to final result – were shorter for nanopore-based tNGS (4.3 days or 2.3 business days) compared to Illumina-based WGS (6.3 days) in our current workflows (Figure 3B). These results demonstrate nanopore-based tNGS can offer comprehensive DR detection before MTBC isolates are available for WGS or culture-based DST.

3.5. tNGS on MTBC-positive cultures

tNGS may also provide additional utility for identifying high-confidence and unknown/novel mutations within MTBC-positive cultures. tNGS was performed on a panel of 55 MTBC-positive cultures, 21 of which were dual-positive for *M. avium* complex. Complete profiles were obtained for 100% isolates, whereas dual-positive cultures yielded either complete (86%) or partial profiles (14%) (Supplementary Table S7). Profiles were in 100% concordance with WGS (Table 2); however, two dual-positive samples did not have WGS available for comparison due to failure to obtain pure MTBC culture. These results demonstrate that tNGS can build comprehensive DR profiles from cultured material, even for mixed cultures that may not meet quality criteria for WGS analysis.

3.6. tNGS SNP-based lineage prediction

In addition to DR profiling, tNGS data can also be used to identify the seven main phylogenetic MTBC lineages to provide supporting data during epidemiological investigations. *In silico* lineage predictions tools often utilize single-nucleotide polymorphisms (SNPs) to classify each lineage, but these SNP catalogs may vary (48). A SNP-based algorithm for lineage prediction was designed using the targets available in the tNGS assay (Figure 4). This algorithm initially relies

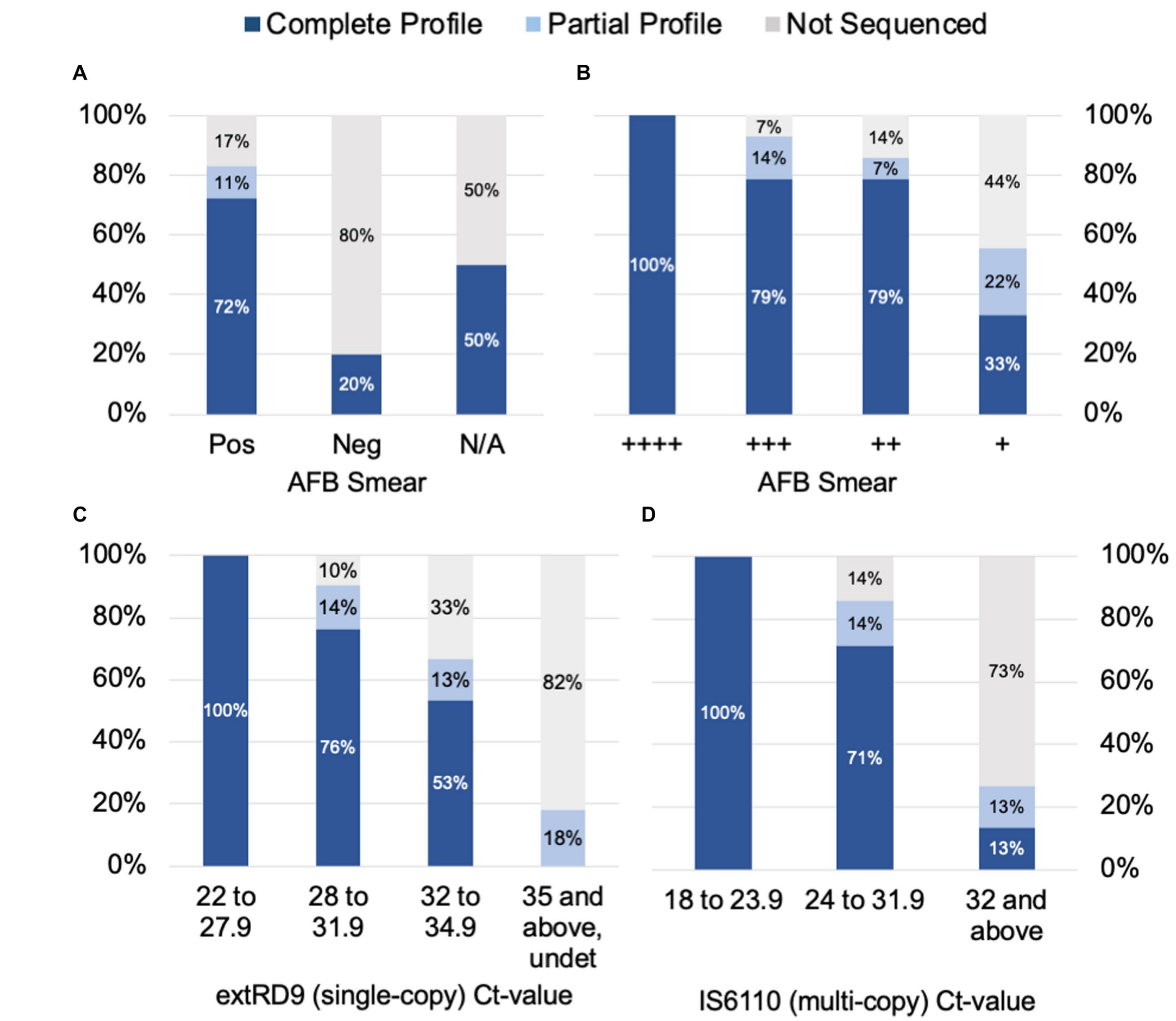


FIGURE 2
tNGS data completeness is arranged by AFB smear **(A)** and **(B)** tNGS data completeness is arranged by AFB smear and real-time PCR values for **(C)** a single-copy target RD9 and **(D)** multi-copy target IS*6110* for MTBC. Profiles are defined as complete (all 13 targets pass QC), partial (≥10 targets pass QC), or not sequenced.

on *gyrB* mutations that distinguish *M. tuberculosis* (*gyrB* 403 GCG mutation) from other *Mycobacterium* species (*gyrB* 403 TCG). For *M. tuberculosis* strains, the algorithm identifies markers for lineage 1 (*gyrB* 291 ATC), lineage 3 (*oxyR-ahpC*, −88 A), and lineage 4 (*katG* 463 CGG). MTBC strains not falling into these categories are classified as likely lineage 2; "likely" reflects the limitation that lineage 2 cannot be distinguished from the rarer lineage 7 with this set of loci. Other members of the MTBC are identified as lineage 5 (*ethA* 124 GAC), lineage 6 or 9 (*inhA* 78 GCG), *M. bovis* or *bovis* BCG (*pncA* 57 GAC), *M. orygis* (*gyrB* 290 GCA), or *M. caprae* (*gyrB* 356 GCG). Strains not fitting these criteria are not assigned with a lineage determination.

These SNP-based lineage predictions were performed on all samples included in this study (primary specimens and cultures) where both tNGS and WGS results were available ($n = 109$). This panel included lineages 1–4 and included one *M. bovis* BCG strain. Comparison of lineages derived *in silico* from tNGS and WGS are shown in Table 3. 98.2% of lineages were correctly identified by tNGS, 0.9% were undetermined due to target failure, and one lineage 4 strain (0.9%) was identified as "likely lineage 2" due to a heterogeneous SNP at *katG* 463. These results show that SNP-based lineage predictions are possible and highly accurate using a small number of loci.

3.7. Fiscal analysis of tNGS

The cost associated with nanopore-based tNGS is detailed in Table 4. The fixed cost per sample includes reagents for extraction ($12.17), PCR ($5.90) and library preparation ($37.21). The cost of gel electrophoresis is not included as this is considered an optional step. Some tNGS costs per sample are dependent on batch size; for example, each tNGS sequencing run requires $25.90 of reagents for flow cell priming, loading, and washing/storing regardless of the number of samples run. Other costs depend on flow cell reusage; we determined costs based on an average of 8 samples per run and up to three flow cell uses. Based on these estimates, the total estimated cost is $78.31 per sample. This analysis also does not include plastic consumables, technician time, instrumentation, or facility overhead as these factors may be facility specific and add to the overall price of the test. The cost

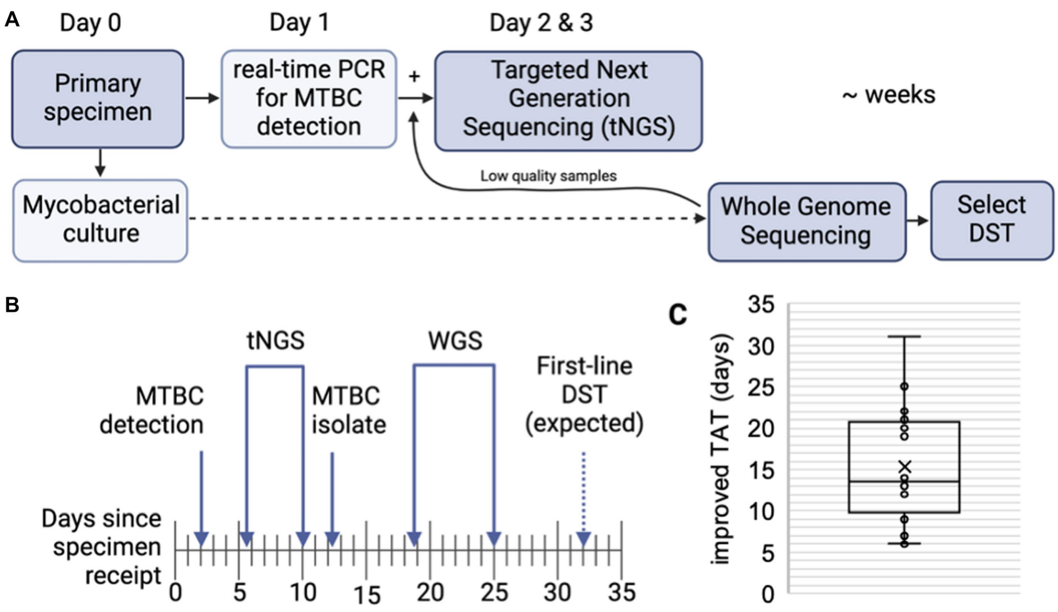


FIGURE 3
Turnaround times for MTBC molecular testing and sequencing. **(A)** MTBC testing algorithm at the Wadsworth Center. Processed specimens are used for mycobacterial culture. Heat killed aliquots are tested for MTBC DNA via real-time PCR and positive specimens are then referred to tNGS. When positive cultures are available, WGS is performed. Phenotypic drug-susceptibility testing (DST) is performed only if unknown/novel mutations or multidrug resistant strains are detected. **(B)** Timeline showing the average number of days required for MTBC DNA detection via real-time PCR, tNGS (from extraction to result), MTBC isolation, and WGS (from extraction to result) (*n* = 16). Note that tNGS and WGS assays are batched weekly and average turnaround times include non-business days (i.e., weekends). Estimated time for first-line DST results are indicated with a dashed arrow. **(C)** Turnaround time (TAT) improvements (in days) of direct tNGS compared to culture-derived WGS.

TABLE 2 Comparison of DR profiles identified by tNGS and WGS performed on MTBC-positive cultures.

		Whole Genome Sequencing (Culture)				
		S	R	MDR	pre-XDR	Not Sequenced
tNGS (Culture)	S	45	0	0	0	2
	R	0	3	0	0	0
	MDR	0	0	3	0	0
	pre-XDR	0	0	0	2	0
	Total	45	3	3	2	2

Profiles are categorized as pan-susceptible (S), mono or poly-resistant (R), multidrug resistant (MDR), and pre-extensively drug resistant (pre-XDR). Not sequenced indicates that a high-quality sample was not available for WGS analysis.

per sample is similar to the cost of high-throughput WGS sequencing currently performed at the Wadsworth Center (49). These analyses indicate that tNGS assays can be cost-effective for implementation in diagnostic/clinical laboratories.

4. Discussion

4.1. tNGS is sensitive, scalable, and reliable for rapid prediction of drug resistance

tNGS represents a sensitive, reliable, and cost-effective method for detecting DR-TB direct from primary specimens in a clinical or diagnostic laboratory setting. This assay accurately identified diverse DR profiles – including MDR and XDR strains – with easier set-up than single-target assays and faster turnaround times than testing performed on cultured MTBC isolates, including WGS and phenotypic DST. This laboratory-developed tNGS assay represents an improvement to our current testing algorithm by offering comprehensive DR profiling shortly after TB diagnosis. Our study revealed a 15 day improvement in turnaround time compared to culture-based WGS, but additional experience will continue to inform tNGS implementation and improve the time to result.

4.2. tNGS assays require careful selection of targets and high-confidence mutations

tNGS assays require careful selection of targets and high confidence resistance mutations (50). Our assay targets full-length loci associated with resistance to first- and second-line MTBC antimicrobials and is consistent with other targeted assays (31–38), with some variation in number and size of loci included. In contrast to molecular beacon and line probe assays which focus just on hot spot regions, the assay described in this study examines full-length genes and promoter regions of many targets to allow for detection of rare and atypical resistance mutations. Although most smear positive specimens yielded complete susceptibility profiles, we found that longer targets (i.e., *embB*, *rpoB*) were more prone to low-coverage or amplification failure, resulting in partial susceptibility profiles. This observation suggests that sensitivity may be improved by splitting larger loci into multiple overlapping amplicons. One additional limitation of tNGS assays is that amplification may fail if strains carry

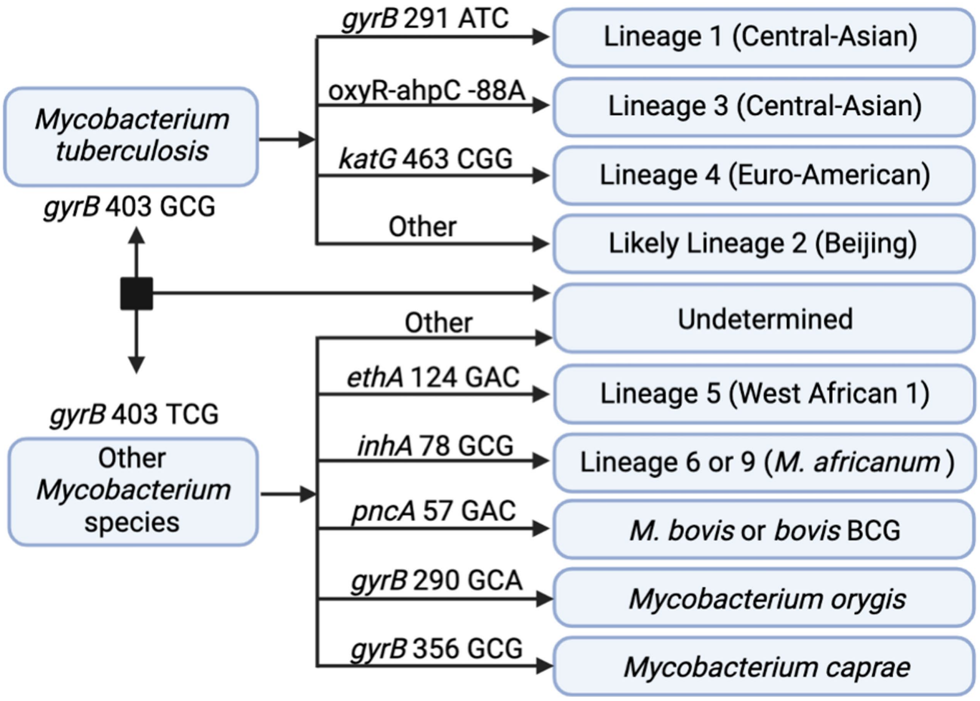

Mycobacterium tuberculosis
gyrB 403 GCG
gyrB 291 ATC
oxyR-ahpC -88A
katG 463 CGG
Other
Lineage 1 (Central-Asian)
Lineage 3 (Central-Asian)
Lineage 4 (Euro-American)
Likely Lineage 2 (Beijing)
Undetermined
gyrB 403 TCG
Other *Mycobacterium* species
Other
ethA 124 GAC
inhA 78 GCG
pncA 57 GAC
gyrB 290 GCA
gyrB 356 GCG
Lineage 5 (West African 1)
Lineage 6 or 9 (*M. africanum*)
M. bovis or *bovis* BCG
Mycobacterium orygis
Mycobacterium caprae

FIGURE 4
In silico SNP-based lineage classifications for MTBC. The SNP-based ID algorithm looks for unique SNPs in *gyrB*, *oxyR-ahpC*, *katG*, *ethA*, *inhA*, and *pncA* in the order shown below. Diagram created with BioRender.com.

TABLE 3 Concordance of *in silico* SNP-based lineage classifications from tNGS and WGS datasets.

		Lineage (WGS)				
		1	2	3	4	BCG
Lineage (tNGS)	Lineage 1	14	0	0	0	0
	Likely Lineage 2	0	31	0	1	0
	Lineage 3	0	0	11	0	0
	Lineage 4	0	0	0	50	0
	M. bovis or BCG	0	0	0	0	1
	Undetermined	1	0	0	0	0
	Total	15	31	11	51	1

TABLE 4 Costs associated with tNGS.

tNGS steps	Total number of samples run on each flow cell			
	1 sample	8 samples	16 samples	24 samples
Extraction	$12.17	$97.36	$194.72	$292.08
PCR	$5.90	$47.22	$94.45	$141.67
Library preparation	$37.21	$297.69	$595.37	$836.40
Nanopore sequencing reagents*	$25.90 (1 run)	$25.90 (1 run)	$51.81 (2 runs)	$77.71 (3 runs)
Nanopore flow cell**	$475.00	$475.00	$475.00	$475.00
Total cost per sample	$556.19	$117.89	$88.21	$78.31

*Includes the cost of reagents for priming, loading, and washing/storing the flow cell. Cost is calculated per run and is independent of number of samples. **Calculations assume up to 8 samples per run and up to three re-uses of the flow cell.

mutations or deletions in primer binding regions, but these undetermined results will be further understood upon reflexing to WGS or culture-based DST in diagnostic testing algorithms (Figure 3A).

The accuracy of sequencing-based predictions for TB DR compared to phenotypic DST have been previously established for WGS (26); however, both tNGS and WGS assays require regular updates to keep pace with the emergence of new DR mutations. In the current Wadsworth Center testing algorithm, isolates with novel mutations undergo DST in order to characterize the potential impact of these mutations. A minimum of three isolates with paired phenotypic DST results or strong supporting literature are required to move novel mutations – initially reported as "unknowns" – to either a neutral or high confidence DR mutation list (26). Laboratories with smaller testing catalogs may refer to the WHO database (51) or other supporting literature to supplement their high confidence DR mutation list. tNGS assays may be updated with additional targets or multiplex pools to keep pace with emerging need, such as genotypic DST predictions for drugs included in the BPaLM/BPaL (bedaquiline, pretomanid, linezolid, moxifloxacin) regimens for treating MDR and XDR infections (52).

4.3. Considerations for implementing Oxford Nanopore sequencing

Special considerations and workflow adaptations are required for using Oxford Nanopore Technologies sequencing platforms. Raw data files from nanopore sequencing devices are basecalled into fastq files that are available for analysis. Newer versions of these basecalling algorithms continually provide better sequencing accuracy and, depending on the algorithm and flow cell version used, these accuracies can approach and even surpass Illumina-based platforms (53). These improvements demand greater processing requirements and thus can create lag times between sequencing and fastq file generation, thus, the use of graphics processing unit (GPU) or Cloud computing resources are highly recommended for basecalling and data post-processing (54, 55). Both the MinION Mk1C and GridION platforms from Oxford Nanopore Technologies were used in this study. While both platforms were able to take advantage of their GPUs for basecalling, we found that the compute power of GridION was able to perform high accuracy basecalling in real-time, enhancing turnaround times compared to the MinION Mk1C. The GridION, however, occupies a larger footprint in the laboratory and is less portable than the Mk1C for applications in resource-limited settings.

Future applications of this technology include detection of heterozygous positions, but this is currently limited by the accuracy of the sequencing data. Newer nanopore chemistries paired with more advanced basecalling algorithms show improved accuracy and potential for heterozygous detection (56). However, these updates to chemistry also necessitate frequent validation and verification. Thus, adoption of commercial products with longevity are critical for clinical implementation and use.

Consistent with other studies (40), we found that nanopore-based tNGS was cost-effective and comparable to current high-throughput WGS assays. Nevertheless, nanopore costs can vary widely depending on batch sizes and flow cell usage. To minimize cost, this validation study successfully obtained accurate tNGS data with re-used flow cells; however, we suggest using unique barcodes for each run to limit potential cross-contamination in clinical testing. Laboratories with lower testing volumes may consider combing multiple targeted assays onto one nanopore flow cell.

In conclusion, this study demonstrates the utility of a clinical tNGS assay as an early detection method for drug resistance direct from MTBC-positive specimens. This particular tNGS assay showed more than a two-week improvement in turnaround time compared to culture and WGS workflows at a similar cost. This method also offers additional utility for cultures that are low quality for WGS analysis due to mixed organisms or low MTBC DNA concentration. Early detection methods are an essential part of TB testing algorithms to ensure that patients are expeditiously placed on appropriate drug treatment regimens.

Data availability statement

Illumina and MinION sequencing datasets are available in BioProject PRJNA766641 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA766641) under the accession numbers listed in Supplementary Table S8.

Author contributions

SM, CS, PL, and KM contributed to conception and design of the study. SM and CS optimized the targeted sequencing methods. SM, CS, KP, and JS collected data. JS identified mutations for *in silico* lineage prediction. PL developed bioinformatic pipelines and wrote sections of the manuscript. SM, PL, TH, MD, VE, M-CR, and KM provided feedback for implementation in the public health laboratory. SM wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research project was supported by a grant from the National Institute of Allergy and Infectious Diseases (5R21AI13985602) and partially supported by NU50CK000516 (Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases in New York State).

Acknowledgments

We acknowledge the Wadsworth Center Applied Genomic Technologies Core Facility for next generation DNA sequencing on Illumina platforms. We also thank Wolfgang Haas, Donna Kohlerschmidt, Susan Wolfe, Amy Chiefari, Alexandra Fiero, Kara Tomlin, Emily McGrath, James Long, and Justine Edwards for their technical contributions. Figures were created with BioRender.com and licensed under the following agreements: MC25HMBAEH, QT258GKZIS, and NC258GL23E.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1206056/full#supplementary-material

References

1. World Health Organization. Annual report of tuberculosis. Annual Global TB Report of WHO (2022) 8:1–68. Available at: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (Accessed March 13, 2023).

2. Torres NMC, Julieth J, Rodríguez JJQ, Andrade PSP, Arriagaid MB, Martins NE. Factors predictive of the success of tuberculosis treatment: a systematic review with meta-analysis. *PLoS One.* (2019) 14:e0226507. doi: 10.1371/journal.pone.0226507

3. Lam E, Nateniyom S, Whitehead S, Anuwatnonthakate A, Monkongdee P, Kanphukiew A, et al. Use of drug-susceptibility testing for Management of Drug-Resistant Tuberculosis, Thailand, 2004–2008. *Emerg Infect Dis.* (2014) 20:408–16. doi: 10.3201/EID2003.130951

4. Xu C, Li R, Shewade HD, Jeyashree K, Ruan Y, Zhang C, et al. Attrition and delays before treatment initiation among patients with MDR-TB in China (2006–13): magnitude and risk factors. *PLoS One.* (2019) 14:e0214943. doi: 10.1371/JOURNAL.PONE.0214943

5. Doulla BE, Squire SB, MacPherson E, Ngadaya ES, Mutayoba BK, Langley I. Routine surveillance for the identification of drug resistant tuberculosis in Tanzania: a cross-sectional study of stakeholders' perceptions. *PLoS One.* (2019) 14:e0212421. doi: 10.1371/JOURNAL.PONE.0212421

6. Zhu J, Bao Z, Xie Y, Werngren J, Hu Y, Davies Forsman L, et al. Additional drug resistance for *Mycobacterium tuberculosis* during turnaround time for drug-susceptibility testing in China: a multicenter observational cohort study. *Int J Infect Dis.* (2021) 108:81–8. doi: 10.1016/J.IJID.2021.04.027

7. van Beek J, Haanperä M, Smit PW, Mentula S, Soini H. Evaluation of whole genome sequencing and software tools for drug susceptibility testing of *Mycobacterium tuberculosis. Clin Microbiol Infect.* (2019) 25:82–6. doi: 10.1016/J.CMI.2018.03.041

8. Van Der Werf MJ, Langendam MW, Huitric E, Manissero D. Multidrug resistance after inappropriate tuberculosis treatment: a meta-analysis. *Eur Respir J.* (2012) 39:1511–9. doi: 10.1183/09031936.00125711

9. Blakemore R, Story E, Helb D, Kop JA, Banada P, Owens MR, et al. Evaluation of the analytical performance of the Xpert MTB/RIF assay. *J Clin Microbiol.* (2010) 48:2495–501. doi: 10.1128/JCM.00128-10

10. Ling DI, Zwerling AA, Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *Eur Respir J.* (2008) 32:1165–74. doi: 10.1183/09031936.00061808

11. Lawn SD, Nicol MP. Xpert® MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. *Future Microbiol.* (2011) 6:1067–82. doi: 10.2217/FMB.11.84

12. Theron G, Zijenah L, Chanda D, Clowes P, Rachow A, Lesosky M, et al. Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial. *Lancet.* (2014) 383:424–35. doi: 10.1016/S0140-6736(13)62073-5

13. Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J, et al. The new Xpert MTB/RIF ultra: improving detection of *Mycobacterium tuberculosis* and resistance to rifampin in an assay suitable for point-of-care testing. *MBio.* (2017) 8:e00812. doi: 10.1128/MBIO.00812-17

14. The use of molecular line probe assays for the detection of resistance to second-line anti-tuberculosis drugs: policy guidance. Available at: https://apps.who.int/iris/handle/10665/246131 (Accessed March 13, 2023).

15. Akalu GT, Tessema B, Petros B. High proportion of RR-TB and mutations conferring RR outside of the RRDR of the *rpoB* gene detected in GeneXpert MTB/RIF assay positive pulmonary tuberculosis cases, in Addis Ababa Ethiopia. *PLoS One.* (2022) 17:e0277145. doi: 10.1371/JOURNAL.PONE.0277145

16. Mvelase NR, Cele LP, Singh R, Naidoo Y, Giandhari J, Wilkinson E, et al. Consequences of *rpoB* mutations missed by the GenoType MTBDRplus assay in a programmatic setting in South Africa. *Afr J Lab Med.* (2023) 12:1975. doi: 10.4102/AJLM.V12I1.1975

17. Fitzgibbon MM, Roycroft E, Sheehan G, Mc Laughlin AM, Quintyne KI, Brabazon E, et al. False detection of rifampicin resistance using Xpert® MTB/RIF ultra assay due to an A451V mutation in *Mycobacterium tuberculosis. JAC Antimicrob Resist.* (2021) 3:dlab101. doi: 10.1093/JACAMR/DLAB101

18. Ajileye A, Alvarez N, Merker M, Walker TM, Akter S, Brown K, et al. Some synonymous and nonsynonymous gyrA mutations in *Mycobacterium tuberculosis* lead to systematic false-positive fluoroquinolone resistance results with the Hain GenoType MTBDRsl assays. *Antimicrob Agents Chemother.* (2017) 61:e02169. doi: 10.1128/AAC.02169-16

19. Getachew E, Adebeta T, Gebrie D, Charlie L, Said B, Assefa DG, et al. Pyrosequencing for diagnosis of multidrug and extensively drug-resistant tuberculosis: a systemic review and meta-analysis. *J Clin Tuberc Other Mycobact Dis.* (2021) 24:100254. doi: 10.1016/J.JCTUBE.2021.100254

20. Halse TA, Edwards J, Cunningham PL, Wolfgang WJ, Dumas NB, Escuyer VE, et al. Combined real-time PCR and *rpoB* gene pyrosequencing for rapid identification of *Mycobacterium tuberculosis* and determination of rifampin resistance directly in clinical specimens. *J Clin Microbiol.* (2010) 48:1182–8. doi: 10.1128/JCM.02149-09

21. Zheng R, Zhu C, Guo Q, Qin L, Wang J, Lu J, et al. Pyrosequencing for rapid detection of tuberculosis resistance in clinical isolates and sputum samples from re-treatment pulmonary tuberculosis patients. *BMC Infect Dis.* (2014) 14:1–8. doi: 10.1186/1471-2334-14-200/TABLES/5

22. Zhao JR, Bai YJ, Zhang QH, Wang Y, Luo M, Yan XJ. Pyrosequencing-based approach for rapid detection of rifampin-resistant *Mycobacterium tuberculosis. Diagn Microbiol Infect Dis.* (2005) 51:135–7. doi: 10.1016/J.DIAGMICROBIO.2004.10.001

23. Mesfin EA, Merker M, Beyene D, Tesfaye A, Shuaib YA, Addise D, et al. Prediction of drug resistance by sanger sequencing of *Mycobacterium tuberculosis* complex strains isolated from multidrug resistant tuberculosis suspect patients in Ethiopia. *PLoS One.* (2022) 17:e0271508. doi: 10.1371/JOURNAL.PONE.0271508

24. Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, Butcher PD. Clinical use of whole genome sequencing for *Mycobacterium tuberculosis. BMC Med.* (2016) 14:46. doi: 10.1186/S12916-016-0598-2

25. Cox H, Goig GA, Salaam-Dreyer Z, Dippenaar A, Reuter A, Mohr-Holland E, et al. Whole-genome sequencing has the potential to improve treatment for rifampicin-resistant tuberculosis in high-burden settings: a retrospective cohort study. *J Clin Microbiol.* (2022) 60:e0236221. doi: 10.1128/jcm.02362-21

26. Shea J, Halse TA, Lapierre P, Shudt M, Kohlerschmidt D, Van Roey P, et al. Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of *Mycobacterium tuberculosis* in New York state. *J Clin Microbiol.* (2017) 55:1871–82. doi: 10.1128/JCM.00298-17

27. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance. Available at: https://www.who.int/publications/i/item/9789240028173 (Accessed March 13, 2023).

28. Park M, Lalvani A, Satta G, Kon OM. Evaluating the clinical impact of routine whole genome sequencing in tuberculosis treatment decisions and the issue of isoniazid mono-resistance. *BMC Infect Dis.* (2022) 22:349. doi: 10.1186/S12879-022-07329-Y

29. Olaru ID, Patel H, Kranzer K, Perera N. Turnaround time of whole genome sequencing for mycobacterial identification and drug susceptibility testing in routine practice. *Clin Microbiol Infect.* (2018) 24:659.e5–7. doi: 10.1016/J.CMI.2017.10.001

30. Tortoli E, Cichero P, Piersimoni C, Simonetti MT, Gesu G, Nista D. Use of BACTEC MGIT 960 for recovery of mycobacteria from clinical specimens: multicenter study. *J Clin Microbiol.* (1999) 37:3578–82. doi: 10.1128/JCM.37.11.3578-3582.1999

31. Leung KSS, Tam KKG, Ng TTL, Lao HY, Shek RCM, Ma OCK, et al. Clinical utility of target amplicon sequencing test for rapid diagnosis of drug-resistant *Mycobacterium tuberculosis* from respiratory specimens. *Front Microbiol.* (2022) 13:974428. doi: 10.3389/FMICB.2022.974428

32. Wu S-H, Xiao Y-X, Hsiao H-C, Jou R. Development and assessment of a novel whole-gene-based targeted next-generation sequencing assay for detecting the susceptibility of *Mycobacterium tuberculosis* to 14 drugs. *Microbiol Spectr.* (2022) 10:e0260522. doi: 10.1128/spectrum.02605-22

33. Gliddon HD, Frampton D, Munsamy V, Heaney J, Pataillot-Meakin T, Nastouli E, et al. A rapid drug resistance genotyping workflow for *Mycobacterium tuberculosis*, using targeted isothermal amplification and nanopore sequencing. *Microbiol Spectr.* (2021) 9:e0061021. doi: 10.1128/SPECTRUM.00610-21

34. Colman RE, Anderson J, Lemmer D, Lehmkuhl E, Georghiou SB, Heaton H, et al. Rapid drug susceptibility testing of drug-resistant *Mycobacterium tuberculosis* isolates directly from clinical samples by use of amplicon sequencing: a proof-of-concept study. *J Clin Microbiol.* (2016) 54:2058–67. doi: 10.1128/JCM.00535-16

35. Barbosa-Amezcua M, Cuevas-Córdoba B, Fresno C, Haase-Hernández JI, Carrillo-Sánchez K, Mata-Rocha M, et al. Rapid identification of drug resistance and phylogeny in *M. tuberculosis*, directly from sputum samples. *Microbiol Spectr.* (2022) 10:e0125222. doi: 10.1128/SPECTRUM.01252-22

36. Jouet A, Gaudin C, Badalato N, Allix-Béguec C, Duthoy S, Ferré A, et al. Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to 13 anti-tuberculous drugs. *Eur Respir J.* (2021) 57:2002338. doi: 10.1183/13993003.02338-2020

37. Bonnet I, Enouf V, Morel F, Ok V, Jaffré J, Jarlier V, et al. A comprehensive evaluation of GeneLEAD VIII DNA platform combined to Deeplex Myc-TB® assay to detect in 8 days drug resistance to 13 Antituberculous drugs and transmission of *Mycobacterium tuberculosis* complex directly from clinical samples. *Front Cell Infect Microbiol.* (2021) 11:707244. doi: 10.3389/FCIMB.2021.707244/FULL

38. Park J, Jang W, Kim M, Kim Y, Shin SY, Park K, et al. Molecular drug resistance profiles of *Mycobacterium tuberculosis* from sputum specimens using ion semiconductor sequencing. *J Microbiol Methods.* (2018) 145:1–6. doi: 10.1016/J.MIMET.2017.12.003

39. Cabibbe AM, Spitaleri A, Battaglia S, Colman RE, Suresh A, Uplekar S, et al. Application of targeted next-generation sequencing assay on a portable sequencing platform for culture-free detection of drug-resistant tuberculosis from clinical samples. *J Clin Microbiol.* (2020) 58:e00632. doi: 10.1128/JCM.00632-20

40. Tafess K, Ng TTL, Lao HY, Leung KSS, Tam KKG, Rajwani R, et al. Targeted-sequencing workflows for comprehensive drug resistance profiling of *Mycobacterium tuberculosis* cultures using two commercial sequencing platforms: comparison of analytical and diagnostic performance, turnaround time, and cost. *Clin Chem.* (2020) 66:809–20. doi: 10.1093/CLINCHEM/HVAA092

41. Tran AC, Halse TA, Escuyer VE, Musser KA. Detection of *Mycobacterium avium* complex DNA directly in clinical respiratory specimens: opportunities for improved turn-around time and cost savings. *Diagn Microbiol Infect Dis*. (2014) 79:43–8. doi: 10.1016/J.DIAGMICROBIO.2014.01.019

42. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res*. (2012) 40:e115. doi: 10.1093/NAR/GKS596

43. Breslauer KJ, Frank R, Blocker H, Marky LA. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*. (1986) 83:3746–50. doi: 10.1073/PNAS.83.11.3746

44. One-pot native barcoding of amplicons v2. Available at: https://www.protocols.io/view/one-pot-native-barcoding-of-amplicons-v2-bp2l6n3rkgqe/v1?step=21 (Accessed March 29, 2023).

45. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*. (2019) 20:1–13. doi: 10.1186/S13059-019-1891-0/FIGURES/2

46. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. (2018) 34:3094–100. doi: 10.1093/BIOINFORMATICS/BTY191

47. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. (2021) 10:1–4. doi: 10.1093/GIGASCIENCE/GIAB008

48. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci Rep*. (2017) 7:46327. doi: 10.1038/SREP46327

49. Rowlinson MC, Musser KA. Current methods and role of next-generation sequencing in the diagnosis of antimicrobial resistance in tuberculosis. *Clin Microbiol Newsl*. (2022) 44:1–12. doi: 10.1016/J.CLINMICNEWS.2021.12.001

50. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in *Mycobacterium tuberculosis* using whole-genome sequencing: progress, promise, and challenges. *Genome Medicine 2019 11:1*. (2019) 11:1–18. doi: 10.1186/S13073-019-0660-8

51. WHO Announces updated definitions of extensively drug-resistant tuberculosis. Available at: https://www.who.int/news/item/27-01-2021-who-announces-updated-definitions-of-extensively-drug-resistant-tuberculosis (Accessed March 22, 2023).

52. Conradie F, Diacon AH, Ngubane N, Howell P, Everitt D, Crook AM, et al. Treatment of highly drug-resistant pulmonary tuberculosis. *N Engl J Med*. (2020) 382:893–902. doi: 10.1056/NEJMOA1901814/SUPPL_FILE/NEJMOA1901814_DATA-SHARING.PDF

53. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genom*. (2023) 9:mgen000910. doi: 10.1099/mgen.0.000910

54. Lin B, Hui J, Mao H. Nanopore technology and its applications in gene sequencing. *Biosensors*. (2021) 11:214. doi: 10.3390/bios11070214

55. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid nanopore genome sequencing in a critical care setting. *N Engl J Med*. (2022) 386:700–2. doi: 10.1056/NEJMc2112090

56. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with nanopore or PacBio sequencing. *Nat Methods*. (2021) 18:165–9. doi: 10.1038/s41592-020-01041-y

frontiers | Frontiers in Cellular and Infection Microbiology

TYPE Brief Research Report
PUBLISHED 04 July 2023
DOI 10.3389/fcimb.2023.1221289

Advantage of precision metagenomics for urinary tract infection diagnostics

Sadia Almas[1], Rob E. Carpenter[1,2], Chase Rowan[1], Vaibhav K. Tamrakar[3,4], Joseph Bishop[1] and Rahul Sharma[1,4*]

[1]Department of Research, Advanta Genetics, Tyler, TX, United States, [2]Soules College of Business, University of Texas at Tyler, Tyler, TX, United States, [3]Divison of Communicable Diseases, ICMR-National Institute of Research in Tribal Health, Jabalpur, India, [4]Department of Research, RetroBioTech LLC, Coppell, TX, United States

Background: Urinary tract infections (UTIs) remain a diagnostic challenge and often promote antibiotic overuse. Despite urine culture being the gold standard for UTI diagnosis, some uropathogens may lead to false-negative or inconclusive results. Although PCR testing is fast and highly sensitive, its diagnostic yield is limited to targeted microorganisms. Metagenomic next-generation sequencing (mNGS) is a hypothesis-free approach with potential of deciphering the urobiome. However, clinically relevant information is often buried in the enormous amount of sequencing data.

Methods: Precision metagenomics (PM) is a hybridization capture-based method with potential of enhanced discovery power and better diagnostic yield without diluting clinically relevant information. We collected 47 urine samples of clinically suspected UTI and in parallel tested each sample by microbial culture, PCR, and PM; then, we comparatively analyzed the results. Next, we phenotypically classified the cumulative microbial population using the Explify® data analysis platform for potential pathogenicity.

Results: Results revealed 100% positive predictive agreement (PPA) with culture results, which identified only 13 different microorganisms, compared to 19 and 62 organisms identified by PCR and PM, respectively. All identified organisms were classified into phenotypic groups (0–3) with increasing pathogenic potential and clinical relevance. This PM can simultaneously quantify and phenotypically classify the organisms readily through bioinformatic platforms like Explify®, essentially providing dissected and quantitative results for timely and accurate empiric UTI treatment.

Conclusion: PM offers potential for building effective diagnostic models beyond usual care testing in complex UTI diseases. Future studies should assess the impact of PM-guided UTI management on clinical outcomes.

KEYWORDS

urinary tract infections, uropathogen, PCR, MNGs, precision metagenomics, next-generation sequencing, UTI management, UTI treatment

OPEN ACCESS

EDITED BY
Varvara K. Kozyreva,
California Department of Public Health,
United States

REVIEWED BY
Xiong Wang,
Minnesota Department of Health,
United States
Candace Cole,
Centers for Disease Control and
Prevention (CDC), United States

*CORRESPONDENCE
Rahul Sharma
rsharma@aalabs.com;
rahuldnadx@gmail.com

RECEIVED 12 May 2023
ACCEPTED 15 June 2023
PUBLISHED 04 July 2023

CITATION
Almas S, Carpenter RE, Rowan C, Tamrakar VK, Bishop J and Sharma R (2023) Advantage of precision metagenomics for urinary tract infection diagnostics. *Front. Cell. Infect. Microbiol.* 13:1221289. doi: 10.3389/fcimb.2023.1221289

1 Introduction

Urinary tract infections (UTIs) are common human illnesses, affecting nearly 50% of people at least once in their life and disproportionately impacting adult women (Sihra et al., 2018). In the United States alone, more than 1 million people suffer from difficult-to-treat or chronic UTIs every year. In clinical settings, UTIs are one of the leading causes of antibiotic prescriptions in adults, which alter urinary tract microbiome and result in antimicrobial resistance—a substantial challenge for public health in recent years (McAdams et al., 2019; Finton et al., 2020). Another important consideration is that complex infectivity can trigger systemic infection with deleterious harm (Neugent et al., 2020; Kaushik et al., 2021). Furthermore, clinical management of UTIs has become more difficult because of resistance to most beta-lactam antibiotics (Rajabnia et al., 2019). No doubt that urinary tract infectivity can lead to costly and unproductive treatment and recurrent disease and trigger undesirable quality of life outcomes (Zhang et al., 2022). And it is likely that much of the UTI diagnostic challenge comes from pairing a matrix and microbiome that is conducive for large numbers of potential pathogens with current limitations in molecular testing (Mouraviev and McDonald, 2018; Lee et al., 2020; Jones-Freeman et al., 2021).

The standard method of uropathogen diagnosis is often microbial culture and susceptibility testing. But because the diagnostic yield of urine culture is frequently influenced by prior antibiotic exposure, poor sensitivity, and difficult-to-culture or uncultivable microorganisms, culturing techniques remain ineffective for up to 50% of symptomatic women (Price et al., 2016). And although polymerase chain reaction (PCR) methods can rapidly detect pathogens directly from clinical samples compared to culturing, including uncultivable microorganisms, PCR methods are limited to amplifying pretargeted species (Smith and Osborn, 2009). There is an unmet need for additional laboratory techniques to timely and accurately detect uropathogens.

In recent years, laboratorians have advanced uropathogen discovery with metagenomic next-generation sequencing (mNGS) (Mouraviev and McDonald, 2018). Unlike PCR, the mNGS approach is target-agnostic and does not require a prior microorganism knowledge. And by sequencing all nucleic acids in a sample, a wide net is cast likely capturing any existing microorganisms, including the urobiome (Figure 1). The application of mNGS has shown promise in various UTI case studies (Li et al., 2020; Duan et al., 2022). But the adoption of mNGS is slow in the clinical laboratory due to the associated costs, expertise, and bioinformatic workflows required (Sharma et al., 2015; Carpenter et al., 2022). Moreover, extracting clinically relevant information can be challenging for target-agnostic approaches—based on rRNA gene amplification and shotgun sequencing after the depletion of host DNA—that promote a high microbiome yield (Price et al., 2021). Alternatively, a hybridization capture-based targeted sequencing approach, also known as precision metagenomics (PM), has potential to bridge the diagnostic gap by providing enhanced discovery power with better diagnostic yield without diluting clinically relevant information (Cariou et al., 2018)—also enabling important uropathogen discovery including fastidious, obligate anaerobic, and non-culturable microorganisms (Stinnett et al., 2021).

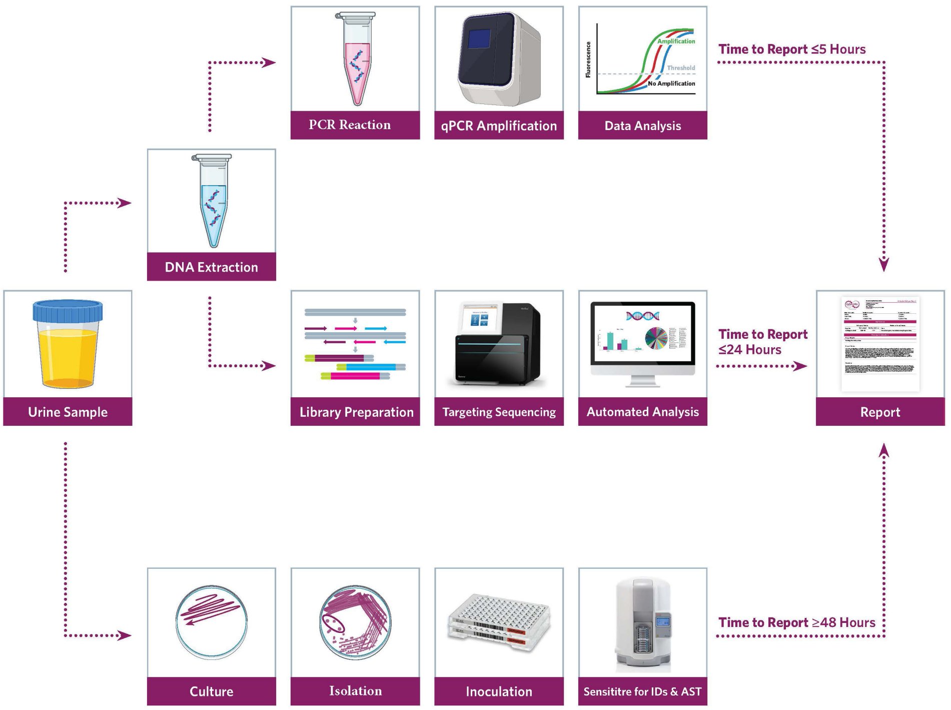


FIGURE 1
Workflow for Urinary Tract Infection (UTI) precision metagenomics analysis.

Accordingly, PM has potential to overcome limitations of both routine and robust UTI testing methods directly from clinical samples.

The purpose of this study is 2-fold. First, we collected 47 urine samples of clinically suspected UTI and in parallel tested each sample by microbial culture, PCR, and PM; then, we comparatively analyzed the results. Second, we phenotypically classified the cumulative microbial population using the Explify® data analysis platform (IDbyDNA) for potential pathogenicity. Last, we discuss the potential clinical benefits of PM for UTI management.

2 Materials and methods

Forty-seven urine samples were collected for routine clinical testing following the standard operating procedures conforming to the rules of aseptic technique and transported to the laboratory. Culture and PCR testing were performed for routine clinical diagnosis. Then, de-identified remnant urine was tested with a hybridization capture-based PM workflow at Advanta Genetics (Tyler, TX, USA).

2.1 Microbial culture

Urine (1 µl) was inoculated onto Spectra UTI biplates (ThermoFisher, Carlsbad, CA, USA) with chromogenic medium for isolation, differentiation, and presumptive pathogen detection. The inoculated medium was incubated for 24 h at 37°C aerobically. After 24 h, biplates were examined for microbial colonies, morphology, and color reactions. Any plates with no growth were incubated for an additional 24 h. Microbial colonies on each side of the biplates were counted, and results were reported as colony-forming units (CFUs) per milliliter of urine. Colony count was reported in log-10 intervals ($<10^4$, 10^4–10^5, or $>10^5$ CFU/ml), where $<10^4$ CFU is considered clinically irrelevant. Preliminary identification was acquired through rapid benchtop testing, chromogenic agar, and Gram stain evaluation. Definitive identification was performed with the Sensititre™ARIS HiQ™ System (ThermoFisher, Carlsbad, CA, USA).

2.2 Nucleic acid extraction

The urine samples were vortexed for a minimum of 10 s to ensure homogeneity before 500 µl of sample was transferred to a 2-ml safe-lock tube (Eppendorf, Hamburg, Germany) containing approximately 100 µl of RNase-free zirconium oxide beads (Next Advance, Inc., Troy, NY, USA) along with 20 µl of proteinase K (Invitrogen, Waltham, MA, USA). Samples were lysed using a TissueLyser (Qiagen Inc., Hilden, Germany) at 30 Hz for 5 min. A 150-µl aliquot of the lysed sample was then combined with 50-µl internal control (IC) in a 96-well plate (Roche, Germany) and loaded into a MagNA Pure 96 System (Roche, Germany) programmed according to the manufacturer's guidelines using the MagNA Pure 96 DNA and Viral NA Small Volume Kit (Roche, Germany) with an elution volume of 100 µl. A synthetic DNA IC was spiked (5 µl) into each sample prior to DNA extraction, and successful extraction was confirmed by positive detection of IC by PCR amplification. Each sample was also spiked with T7 bacteriophage DNA (Microbiologics, St. Cloud, MN, USA), delivering a final concentration of 1.2×10^7 PFU/ml of the sample. Copies of T7 were used for computing the absolute concentration of the target copies detected by PM.

2.3 qRT-PCR testing

Each sample was tested for 28 uropathogens (24 bacteria and 4 fungi) (S-1) using commercially available predesigned PCR reaction mixtures (Scienetix, Tyler, TX, USA). Briefly, 2.5 µl of extracted DNA was added to a 7.5-µl reaction mixture containing the microbe-specific primer pairs and TaqMan probes. Triplex PCR reactions were performed on the Light Cycler® instrument (Roche, Germany) in a 384-well plate with the thermal cycling program set to initial denaturation at 95°C for 3 min followed by 40 cycles of amplification at 95°C for 5 s and 60°C for 30 s. An amplification control (AC) containing the target-specific template DNA for each microbe was tested as a positive control (PC), while molecular grade water was tested as no template control (NTC). The quantitative cycle threshold (Ct) value of ≤35, when accompanied by the sigmoid amplification curve, was considered positive for the qualitative detection of the targeted organism.

2.4 Precision metagenomics

2.4.1 Library preparation and sequencing

Sequencing libraries were prepared using IDbyDNA Urinary Pathogen ID/AMR Panel (UPIP) protocol (Illumina Inc, San Diego, CA, USA). An aliquot of the DNA used for PCR testing was used for library preparation and sequencing. Libraries were constructed by DNA tegmentation and adapter ligation using the Illumina® DNA prep with the enrichment tegmentation kit (Illumina Inc, San Diego, CA, USA). Indexed libraries were enriched for microbial content by hybridization capture of relevant genomic regions of 135 bacteria, 35 viruses, 14 fungi, and 7 parasites (S-1). Indexed libraries were pooled in triplicate and hybridized with the UTI Pathogen ID-AMR probes (Illumina Inc, San Diego, CA, USA) at 95°C for 1 min, followed by 94°C to 58°C with 2-min hold at each 2°C temperature decrement, and 90-min hold at 58°C. Captured libraries were amplified for 18 cycles and cleaned using AmPureXP (Beckman Coulter, Pasadena, CA, USA) beads.

Ten-fold serial dilutions of ZymoBIOMICS (Cat # D6300, Zymo Research, Irvine, CA, USA) community standard was used as a training set to determine reporting thresholds based on sequence data. Genomic coverage, median depth, and reads per kilobase per million mapped (RPKM) resulting in ≥90% accurate detection of known microorganisms were recognized as cutoffs for accepting positive microbe detection. A ZymoBIOMICS community sample and a urine conditioning buffer sample were also processed as PC and negative control (NC), respectively, with each batch of library preparation and sequencing. Libraries were quantified using a Qubit 2.0 fluorometer (Invitrogen, Waltham, MA, USA), and fragment sizes were analyzed in Agilent 5200 Fragment Analyzer (Agilent, Austin, TX, USA). The libraries were then pooled to an equimolar concentration and normalized to 1-nM concentration. The final library pool was denatured and neutralized with 0.1 N NaOH and 200 mM Tris-HCl (pH 8), respectively. The denatured libraries were further diluted to a loading concentration of 2 pM. Dual indexed paired-end sequencing with 75-bp read length was done using the HO flow cell (150 cycles) on the Illumina MiniSeq® instrument.

2.4.2 Explify® bioinformatic analysis

Sequencing data were analyzed with the Explify® UPIP data analysis solution. MiniSeq® run parameters were uploaded on the Explify® portal, and the corresponding run folders containing the binary base call (BCL) sequencing files were shared via a local host. Sequencing data were de-multiplexed using sample-specific barcodes. Samples passing the predefined QC requirements were analyzed. Predefined targets included in the ZymoBIOMICS community were correctly identified from a minimum of 0.5 million reads; meaning, only samples with ≥0.5 million reads were considered for a comprehensive microbial profile. Individual sample results were automatically reported by JavaScript Object Notation format containing the quantitative identification of microorganisms in each sample. Identified organisms were auto classified into phenotypic categories based on the microbe's potential pathogenicity. Group-0 microorganisms were considered common contaminants or healthy microflora; group-1 microorganisms were phenotypically classified as part of the normal flora, colonizers, or contaminants; group-2 microorganisms were phenotypically classified as frequently associated with UTI disease; and group-3 microorganisms were phenotypically classified as routinely pathogenic for UTI disease.

3 Results

Comparative method analysis demonstrated dissimilar diagnostic yield (Tables 1–3) with PM identifying polymicrobial infection in 46/47 (98%) samples compared to PCR 39/47 (83%) and urine culture 33/47 (70%). However, PM had 92% positive predictive agreement (PPA) with culture and 95% PPA with PCR. Urine culture isolated 13 different microorganisms, PCR amplified 19 microorganisms, and PM identified 62 distinct microorganisms. Importantly, PM demonstrated positive results in 13 no growth urine culture biplates resulting in the discovery of 58 additional microorganisms (Figure 2).

3.1 Microbial culture

After 48 h, samples were considered positive if microbial colonies were visible on the primary biplate. Results showed microbial growth in 33/47 (70%) samples, with four samples demonstrating differentiated polymicrobial colonies. *Escherichia coli* was the most commonly isolated organism, 14/33 (30%). Other isolated organisms included *Enterococcus faecalis*, 7/33 (15%); *Proteus mirabilis*, 3/33 (6%); *Citrobacter freundii*, 3/33

TABLE 1 Prevalence of phenotypic group-0 and group-1 microorganisms detected by urine culture, PCR, and PM of suspected UTI cases (n = 47).

Microorganism	Type	Culture (+)	PCR (+)	NGS (+)
Human papillomavirus type 51, 55/44 56 and 68 (HPV; High-risk)	Virus	0	No target	6
Trichomonas vaginalis	parasite	0	No target	1
Actinobaculum massiliense	Bacteria	0	No target	3
Alloscardovia omnicolens	Bacteria	0	No target	1
Corynebacterium aurimucosum	Bacteria	0	No target	3
Corynebacterium coyleae	Bacteria	0	No target	1
Epstein–Barr virus (EBV)	Virus	0	No target	1
Facklamia hominis	Bacteria	0	No target	10
JC polyomavirus	Virus	0	No target	13
Lactobacillus species	Bacteria	2	No target	0
Mobiluncus curtisii	Bacteria	0	No target	4
Peptostreptococcus anaerobius	Bacteria	0	No target	2
Porphyromonas asaccharolytica	Bacteria	0	No target	6
Propionimicrobium lymphophilum	Bacteria	0	No target	17
Rothia kristinae	Bacteria	0	No target	1

TABLE 2 Prevalence of phenotypic group-2 microorganisms detected by urine culture, PCR, and PM of suspected UTI cases (n = 47).

Microorganism	Type	Culture (+)	PCR (+)	NGS (+)
Acinetobacter pittii	Bacteria	0	No target	2
Actinotignum sanguinis (Actinobaculum schaalii)	Bacteria	0	7	9
Aerococcus christensenii	Bacteria	0	No target	1
Aerococcus lactolyticus	Bacteria	0	No target	3
Atopobium vaginae	Bacteria	0	No target	2
Bacteroides fragilis	Bacteria	0	3	3
Bifidobacterium breve	Bacteria	0	No target	3
BK polyomavirus	Virus	0	No target	3
Corynebacterium glucuronolyticum	Bacteria	0	No target	1
Finegoldia magna (Peptostreptococcus magnus)	Bacteria	0	No target	4
Human adenovirus B	Virus	0	No target	1
Oligella urethralis	Bacteria	0	No target	1
Prevotella bivia	Bacteria	0	4	0
Prevotella timonensis	Bacteria	0	No target	11
Providencia stuartii	Bacteria	0	No target	2
Staphylococcus epidermidis	Bacteria	1	No target	5
Staphylococcus haemolyticus	Bacteria	0	No target	2
Staphylococcus hominis	Bacteria	0	No target	2
Staphylococcus simulans	Bacteria	0	No target	3
Staphylococcus warneri	Bacteria	0	No target	1
Streptococcus anginosus	Bacteria	0	No target	6
Streptococcus constellatus	Bacteria	0	No target	1
Streptococcus intermedius	Bacteria	0	No target	1
Ureaplasma parvum	Bacteria	0	No target	1

(6%); *Enterobacter cloacae*, 2/33 (4%); *Klebsiella oxytoca*, 2/33 (4%); while *Staphylococcus aureus*, *Acinetobacter baumannii*, *Enterobacter gergoviae*, *Enterococcus faecium*, *Klebsiella pneumoniae*, and *Staphylococcus epidermidis* were all respectively identified in 1/33 (2%) samples. Two samples (4%) were culture-positive for the *Lactobacillus* genus, and the remaining 14 (30%) were culture-negative—no growth was observed after 48 h of incubation.

Phenotypic classification results placed *Lactobacillus* in group-1 because it is part of the normal vaginal flora and is often considered a contaminant when cultured from urine specimens (Das Purkayastha, 2020; Das Purkayastha et al., 2020). The sample positive for *S. epidermidis* was phenotypically classified in group-2; although *S. epidermidis* is often considered a urine contaminant, it can also be linked to nosocomial infection (Otto, 2009) and cause UTIs in children (Hall and Snitzer, 1994). The remaining positive cultures were phenotypically classified in group-3—common uropathogens often with correlated etiology for UTIs.

3.2 qRT-PCR

Among the 47 samples tested, 39/47 (83%) were positive for ≥1 microbe, while 8/47 (17%) were PCR-negative. Of the 28 microorganisms targeted by PCR, 19/28 (68%) amplified with Ct ≤35 were considered positive. Noticeably abundant were *E. coli* 13/39 (33%) and *Enterobacter cloacae* 13/39 (33%). This was followed by *E. faecalis*, 12/39 (31%); *E. faecium*, 9/39 (23%); *K. pneumoniae*, 9/39 (23%); *Actinobaculum schaali*, 7/39 (18%); *Morganella morganii*, 6/39 (15%); *Pseudomonas aeruginosa*, 6/39 (15%); *Prevotella bivia*, 4/39 (10%); *Bacteroides fragilis*, 3/39 (8%); *K. oxytoca*, 3/39 (8%); *Streptococcus agalactiae*, 3/39 (8%); *Candida albicans*, 2/39 (5%); *Candida glabrata*, 2/39 (5%); *Candida parapsilosis*, 2/39 (5%); and *Citrobacter freundii*; *K. aerogenes* and *Staphylococcus aureus* were positive in 1/39 (3%) of samples.

No microorganisms were classified into phenotypic group-0 or group-1. *A. schaali* and *B. fragilis* were classified in phenotypic group-2—detected in 14% (7/47) and 6% (3/47) of samples,

TABLE 3 Prevalence of phenotypic group-3 microorganisms detected by urine culture, PCR, and PM of suspected UTI cases (n = 47).

Microorganism	Type	Culture (+)	PCR (+)	NGS (+)
Acinetobacter baumannii	Bacteria	1	Negative	0
Aerococcus urinae	Bacteria	0	Negative	6
Candida albicans	Fungi	0	2	0
Candida glabrata	Fungi	0	2	1
Candida parapsilosis	Fungi	0	2	0
Citrobacter freundii	Bacteria	3	1	4
Corynebacterium pseudogenitalium	Bacteria	0	No target	4
Corynebacterium urealyticum	Bacteria	0	No target	3
Enterobacter cloacae complex	Bacteria	2	13	6
Enterococcus faecalis	Bacteria	7	12	12
Enterococcus faecium	Bacteria	1	9	2
Enterococcus raffinosus	Bacteria	0	No target	1
Escherichia coli	Bacteria	14	13	22
Klebsiella aerogenes	Bacteria	1	1	1
Klebsiella. oxytoca	Bacteria	2	3	10
Klebsiella pneumoniae	Bacteria	1	9	7
Klebsiella quasipneumoniae	Bacteria	0	No target	2
Klebsiella variicola	Bacteria	0	No target	2
Morganella morganii	Bacteria	0	6	3
Proteus mirabilis	Bacteria	3	7	5
Pseudomonas aeruginosa	Bacteria	0	6	4
Salmonella enterica	Bacteria	0	No target	3
Serratia marcescens	Bacteria	0	No target	1
Staphylococcus aureus	Bacteria	1	1	1
Streptococcus agalactiae	Bacteria	0	3	2

respectively. Seventeen likely pathogenic microorganisms were phenotypically classified in group-3.

3.3 Precision metagenomics

Ten-fold serial dilutions of the ZymoBIOMICS microbial community were tested in triplicate, and only dilutions with ≥0.5 million total reads resulted in the accurate detection of all targets included in the control. Thus, the minimum yield of 0.5 million reads/sample was applied as a cutoff for further analysis. Furthermore, ≥25% organism target coverage, median depth of ≥1X, and RPKM >10 were identified as acceptance criteria for reporting individual organism (S-2).

Only one sample failed to yield minimum 0.5M reads, and the hybridization capture-based approach was significant for resulting in 46/47 (98%) positive samples. Categorically, PM detected 62 distinct species consisting of 52 bacteria, 8 viruses, 1 fungus, and 1 parasite. The top 5 pathogenic bacteria were *E. coli*, 22/46 (48%); *E. faecalis*, 12/46 (26%); *K. oxytoca*, 10/46 (22%); *K. pneumoniae*, 7/46 (15%); and *Aerococcus urinae*, 6/46 (13%). The positive rate for virus detection was 23/46 (50%)—JC polyomavirus 13/46 (28%) was recognized as the most commonly detected virus. *C. glabrata* 1/46 (2%) and *Trichomonas vaginalis* 1/46 (2%) were the fungal and parasite species detected.

Bioinformatic analysis classified the 62 microorganisms in phenotypic groups. Human papillomavirus (serotypes 51, 55, 56, and 68) and *T. vaginalis* were classified in group-0. Considered common urine contaminants but rarely pathogenic for UTI, 12 organisms were classified in group-1. Considered more frequently associated with UTI disease, group-2 accounted for 23/62 (37%) microorganisms, while 22/62 (35%) of microorganisms were classified in group-3 as likely pathogenic for UTI.

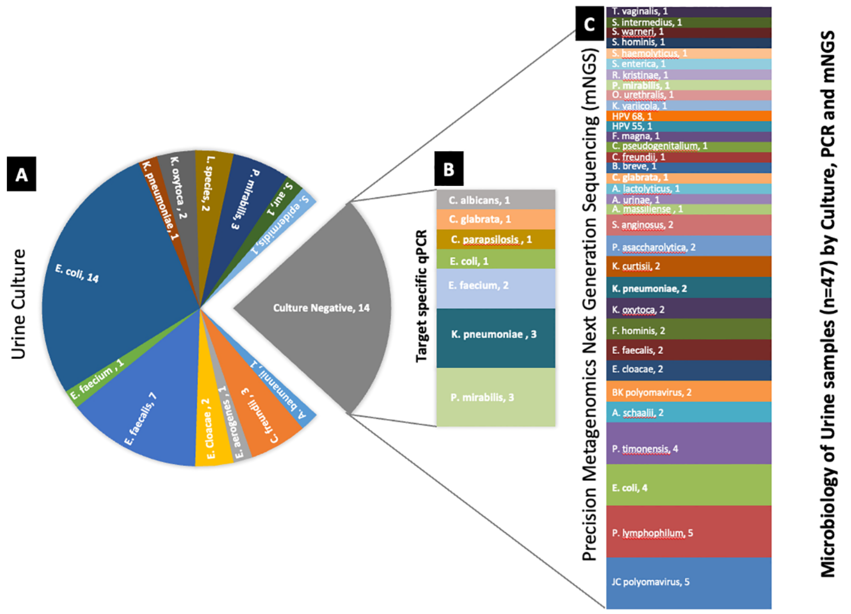


FIGURE 2
Deciphering the microbiology of the culture-negative urine samples using target-specific PCR and PM. **(A)** Microorganisms identified by urine culture. **(B)** Differential microbial profile of the culture-negative samples by target-specific PCR panel. **(C)** Microbial profile of the culture-negative samples detected by PM. The number after the organism's name denotes the number of samples found positive for the organism.

4 Discussion

Several laboratory methods have been developed for the diagnosis of UTIs. The most common include microbiological culture and various nucleic acid amplification techniques. However, these usual care approaches have limitations in the management of UTI and often result in empiric therapy challenges (Cai et al., 2012; Boyanova et al., 2022). Accordingly, this study sought to compare UTI samples for diagnostic yield between urine culture, PCR, and PM. Although the results demonstrate that PM shows promise for guiding better urinary tract diagnostics and therapeutic management of UTI, interpretation of urobiome pathogenicity remains methodologically challenging in the laboratory (Perez-Carrasco et al., 2021). We provide further evidence of this by discussing comparative results and phenotypic classification of the microorganisms detected in this study.

4.1 Comparative results

Culture growth isolated *Lactobacillus* in 2/47 samples; the genus is considered essential for maintaining urinary tract symbiotic microflora (Stapleton and Stamm, 1997). But the genus was not probed by PCR or PM despite research showing quantitative detection informative for microflora imbalance and UTI (Martinez et al., 2014). There were noted concerns between the methods with closely related species. Although the exact probe sequences used in the PM kit were not available, the observed discrepancies indicate it may be that the hybridization capture-based approach is less discriminatory with certain microorganisms. For example, PM failed to differentiate *Prevotella timonensis* from *Prevotella bivia*. This was evident when *P. timonensis* was identified in 11/47 PM tested samples. But all 47 PM samples were negative for *P. bivia* by PM despite PCR amplifying 4/11 samples for *P. bivia*. While *P. timonensis* was not probed by PCR, results suggest specificity concerns for PM false negative for *P. bivia* and likely false positive for *P. timonensis*. Furthermore, *C. albicans* and *C. parapsilosis* were amplified by PCR but not detected by PM. But because the RPKM for these fungal species was below prespecified thresholds for reporting PCR and PM, results were excluded (S-2). We also noted discordance among culture, PCR, and PM results for one *E. coli* sample when it was isolated in urine culture and detected by PM but did not amplify by PCR, suggesting that strain specificity of PCR primers warrants consideration. In another example, one urine culture isolated *A. baumannii*, but it was not detected by PCR or PM. While sequencing techniques of this Gram-negative bacterium have shown genotyping advantages (Alshahni et al., 2015), there are noted discordances in the literature with various laboratory methods, particularly for bacteremia (Pailhoriès et al., 2018). However, of the two culture-positive samples for *A. baumannii*, one of the samples probed PM positive for *Acinetobacter pittii*, which is part of the *A. baumannii* complex, suggesting higher differentiation potential by PM. Enhanced detection was also noted when *A. urinae*—an emerging pathogen causing UTI in older adults (Higgins and Garg, 2017)—was identified in six PM samples but absent in culture and PCR. These noted differentiation challenges (Figure 3) are also seen in similar studies comparing laboratory methods for UTI pathogenicity (Chen et al., 2020).

Implications of organism classification were considered important for clinical relevance (Figure 4). Our downstream analysis utilized the Explify® bioinformatic platform that

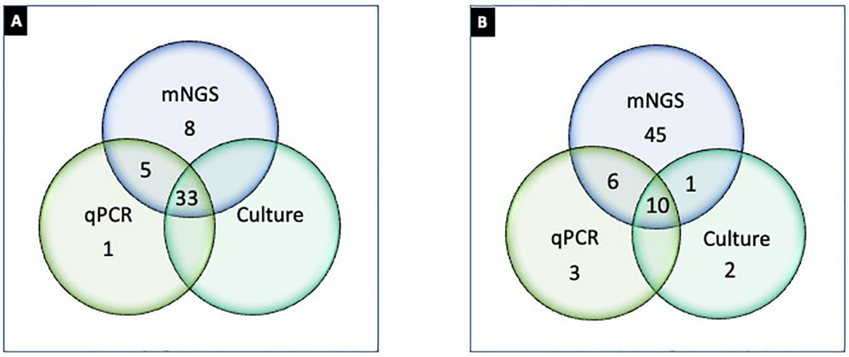


FIGURE 3
Differential detection of microorganisms by urine culture, PCR, and PM in UTI samples (n = 47). **(A)** Number of samples positive (≥1 organism detected) by culture (Greub and Raoult, 2002), PCR (Reploeg et al., 2001), and PM (Goldberg et al., 2015); **(B)** microorganisms detected by culture (Duan et al., 2022), PCR (Stinnett et al., 2021), and PM (62). *Note:* **(A)** represents 33 samples concordantly positive by culture, PCR, and NGS; one sample only positive by PCR; eight samples only positive by PM; five samples positive by PCR and NGS but not by culture. **(B)** represents two microorganisms that were exclusively detected by urine culture, three were exclusively detected by PCR, and 45 were exclusively detected by PM. Ten microorganisms were detected concordantly by culture, PCR, and PM. One microbe was detected by PM and culture, and six microorganisms were detected by PCR and PM.

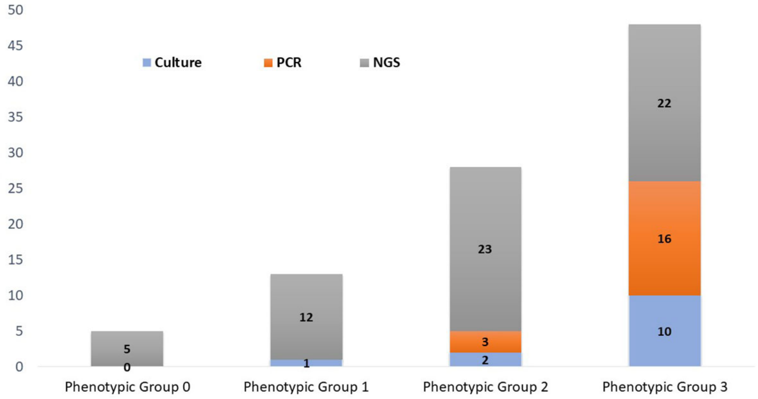


FIGURE 4
Phenotypic classification of microorganisms detected by urine culture, PCR, and PM. Phenotypic group 1: Microorganisms rarely associated with urinary tract infections and may frequently represent normal flora, colonizers, or contaminants. Phenotypic group-2: Microorganisms infrequently associated with urinary tract infections and may frequently represent part of the normal flora, colonizers, or contaminants. Phenotypic group-3: Microorganisms commonly associated with urinary tract infections but may also represent part of the normal flora, colonizers, or contaminants.

classified each microbe into phenotypic groups 0–3[1]—applying an escalating order of potential pathogenicity. Human papillomavirus and *T. vaginalis* were the only microorganisms classified in phenotypic group-0 because both are common etiological agents of sexually transmitted infection, not UTI. Microorganisms classified in phenotypic group-1 are frequently considered part of the normal flora but with the potential for associated UTI diseases in certain clinical manifestations. For example, *Actinobaculum massiliense*, *Corynebacterium aurimucosum*, *Corynebacterium coyleae*, *Peptostreptococcus anaerobius*, and *Propionimicrobium lymphophilum* are part of the commensal microflora of the skin, urethra, mucous membranes, and genital tract but are also reported pathogenic for mild to severe UTI complications (Greub and Raoult, 2002). Likewise, *C. coyleae* can be considered as contamination or normal flora if co-isolated with *E. faecalis* or *E. coli*, but the microbe can be infective if isolated as monoculture (Sokol-Leszczynska et al., 2019). As an example, *C. coyleae* infection in a polycystic kidney disease patient led to bilateral nephrectomy, suggesting that *Corynebacterium* is an emerging pathogen with the potential for complicated UTIs (Barberis et al., 2018). Two viruses were detected by PM and classified in phenotypic group-1—Epstein–Barr virus and JC polyomavirus—both unlikely to be

[1] The Explify® bioinformatic analysis was limited to qualitative detection because of the comparative methods (culture and qualitative PCR) used in the study. The Explify® platform is capable of reporting absolute abundance (organism/ml) of organisms in clinical specimens derived from the RPKM value of a known quantity of spiked T7 Phase. Quantitative analysis was beyond this study. See https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/explify-upip-data-analysis.html

detected by routine microbiology or PCR. The viruses are linked to persistent bladder inflammation and progressive multifocal leukoencephalopathy, respectively (Jhang et al., 2018).

Phenotypic group-2 microorganisms are infrequently associated with UTIs and may frequently represent normal flora, colonizers, or contaminants. However, these microorganisms are routinely cited for UTI pathogenic potential. Of specific note, PM identified 23 microorganisms that were classified in phenotypic group-2, whereas concordance in culture and PCR was only 1/23 and 2/23, respectively. Curiously, *S. epidermidis* was the only *Staphylococcus* species isolated by urine culture (1/47). However, *S. epidermidis* was detected in five PM samples. PM also identified four additional *Staphylococcus* species, three of which were classified in phenotypic group-2. Importantly, these cocci species are emerging opportunistic uropathogens (Kanuparthy et al., 2020). Also classified in phenotypic group-2, BK polyomavirus was PM-positive in 3/47 samples. Although human polyomaviruses are common in the general population, their presence in immunocompromised or immunosuppressed individuals may cause several clinical manifestations. For instance, BK nephropathy is complicit for up to 80% of kidney transplant failures within 2 years, especially if untimely diagnosed (Ambalathingal et al., 2017). Due to its severe consequences, the timely and accurate detection of BK polyomavirus is critical (Reploeg et al., 2001; Myint et al., 2022).

Phenotypic group-3 microorganisms are generally considered uropathogens and rarely present as commensals or contaminants. Phenotypic group-3 microbe detection by method consists of 11 culture, 16 PCR, and 25 PM microorganisms. However, six crucial uropathogens (*Corynebacterium pseudogenitalium*, *Corynebacterium urealyticum*, *Enterococcus raffinosus*, *Klebsiella quasipneumoniae*, *Klebsiella variicola*, and *Salmonella enterica*) were not probed by PCR nor represented by culture growth. The *Enterobacter* species, particularly *E. coli*, were the most prevalent UTI pathogens identified in this study. Even though *Enterobacter* were commonly diagnosed by urine culture and PCR, PM provided greater speciation, important for therapeutic management (Mezzatesta et al., 2012).

Despite some explicable differentiation challenges, PM appears to have greater discovery power in deciphering the microbiome of urine samples—identifying 35 bacterial species not isolated by urine culture or detected by PCR (of these 35 bacterial species, 33 were absent target-specific PCR probes). Importantly, PM exclusively identified eight bacterial species classified in phenotypic group-3 that were undetected by both culture and PCR, including *A. urinae*, *K. quasipneumoniae*, *K. variicola*, *C. urealyticum*, *C. pseudogenitalium*, *S. enterica*, *Serratia marcescens*, and *E. raffinosus.* We note that 6/8 of these bacterial species were not probed by PCR nor do they appear to be common targets for PCR detection. Although *A. urinae* and *S. marcescens* were probed, they failed to amplify, whereas PM detected *A. urinae* in six samples and *S. marcescens* in one sample. Moreover, PM exclusively identified eight viral strains that went undetected by culture or PCR. And although considered a common cause of sexually transmitted infection, *T. vaginalis* was the one parasite detected among 47 samples—and only detected by PM. This is important because many microbiology laboratories are ill-equipped to isolate and identify UTI viruses and parasites (Szlachta-McGinn et al., 2022).

Mixed cultures in clinical microbiology laboratories are often considered possible periurethral or vaginal contamination (Szlachta-McGinn et al., 2022). However, UTIs caused by polymicrobial flora are common (Detweiler et al., 2015). In this study, PM identified coinfection in 41/47 samples—urine culture was polymicrobial in 36 samples. And although PCR has potential for identifying coinfection, recognition is restricted to probed targets (Wojno et al., 2020). For this and other reasons, the hybridization capture-based approach used in this study has potential to improve UTI (co)pathogenic discovery and management.

Despite numerous advantages of PM, the approach still has some constraints before full adoption in the clinical laboratory. Implementing PM in the clinical laboratory is expensive and time-consuming and requires high-level expertise. Furthermore, production workflows may benefit from multiplexity optimization to reduce the cost (Carpenter et al., 2023). Moreover, the extensive differentiation power of PM makes distinguishing between pathogenic and commensal microflora challenging, particularly for less studied and emerging pathogens. Bioinformatic platforms like Explify® are showing promise and will likely improve as more clinically correlated sequencing data emerge. For example, the Explify® platform is capable of reporting absolute abundance (organisms per milliliter) of organisms in clinical specimens, providing clinicians a comprehensive report containing quantitative values of each identified organism in a patient sample—including each organism-associated AMR marker. And although the qualitative detection and AMR marker analysis were beyond the scope of this study, such offerings can help clinicians make a better medical decision for symptomatic patients with clinical UTI symptoms. Yet, despite the paradigm shift to genotyping for diagnosing infectious disease, adoption of the technology by clinicians appears slow (Goldberg et al., 2015). Although sequencing cost and turnaround time are continuously declining, several technical, clinical, and regulatory challenges still delay the broader acceptance of PM in infectious disease management (Goldberg et al., 2015).

However, even in its current qualitative format, this study demonstrates that PM has potential to influence diagnostic specificity and improve public health decisions. As technology continues to advance our understanding of the etiological relationship between microorganisms and their hosts, PM has the potential to bridge the gap between microbial research and diagnostic microbiology for UTIs and other infectious diseases (Almas et al., 2023). Further studies are warranted to demonstrate the financial incentives for accurate and timely diagnosis leading to prompter patient recovery and savings in treatment costs. Moreover, clinical utility studies will likely drive adequate reimbursement of PM for wider adoption in clinical practice.

5 Conclusion

History shows that usual care testing is not the diagnostic solution for recurrent and complex UTI. This study supports that PM offers prospects to bridge the UTI diagnostic gap. This approach allows a workflow where laboratorians can qualify, quantify, and phenotypically classify pathogenicity more readily through bioinformatic platforms like Explify®, essentially providing dissected results across a broad array of input types and quantities for timely and accurate empiric UTI treatment. Moreover, PM offers potential for building effective diagnostic models beyond usual care testing in complex and coinfected UTI diseases. Future studies should assess the impact of PM-guided UTI management on clinical outcomes.

Data availability statement

The data presented in the study are deposited in the NCBI-Sequence Read Archive (SRA), accession number PRJNA986135.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Conceptualization, RC and RS. Methodology, SA and RS. Executed the experiments, CR, JB. Data curation, SA, and VT. Writing—original draft preparation, SA and RS. Writing—review and editing, RC. Supervision, RS. Project administration, RS. All authors have reviewed and approved manuscript for publication.

Conflict of interest

Author RS was employed by the company RetroBioTech LLC. Author RC and RS have a commercial interest in Scienetix LLC.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2023.1221289/full#supplementary-material.

References

Almas, S., Carpenter, R. E., Singh, A., Rowan, C., Tamrakar, V. K., and Sharma, R. (2023). Deciphering microbiota of acute upper respiratory infections: a comparative analysis of PCR and mNGS methods for lower respiratory trafficking potential. *Adv. Respir. Med.* 91 (1), 49–65. doi: 10.3390/arm91010006

Alshahni, M. M., Asahara, M., Kawakami, S., Fujisaki, R., Matsunaga, N., Furukawa, T., et al. (2015). Genotyping of acinetobacter baumannii strains isolated at a Japanese hospital over five years using targeted next-generation sequencing. *J. Infect. Chemother.* 21 (7), 512–515. doi: 10.1016/j.jiac.2015.03.009

Ambalathingal, G. R., Francis, R. S., Smyth, M. J., Smith, C., and Khanna, R. (2017). BK polyomavirus: clinical aspects, immune regulation, and emerging therapies. *Clin. Microbiol. Rev.* 30 (2), 503–528. doi: 10.1128/CMR.00074-16

Barberis, C. M., Montalvo, E., Imas, S., Traglia, G., Almuzara, M. N., Rodriguez, C. H., et al. (2018). Total nephrectomy following. *JMM Case Rep.* 5 (9), e005149. doi: 10.1099/jmmcr.0.005149

Boyanova, L., Marteva-Proevska, Y., Markovska, R., Yordanov, D., and Gergova, R. (2022). Urinary tract infections: should we think about the anaerobic cocci? *Anaerobe* 77, 102509. doi: 10.1016/j.anaerobe.2021.102509

Cai, T., Mazzoli, S., Mondaini, N., Meacci, F., Nesi, G., D'Elia, C., et al. (2012). The role of asymptomatic bacteriuria in young women with recurrent urinary tract infections: to treat or not to treat? *Clin. Infect. Dis.* 55 (6), 771–777. doi: 10.1093/cid/cis534

Cariou, M., Ribière, C., Morlière, S., Gauthier, J. P., Simon, J. C., Peyret, P., et al. (2018). Comparing 16S rDNA amplicon sequencing and hybridization capture for pea aphid microbiota diversity analysis. *BMC Res. Notes.* 11 (1), 461. doi: 10.1186/s13104-018-3559-3

Carpenter, R. E., Tamrakar, V. K., Almas, S., Sharma, A., Rowan, C., and Sharma, R. (2023). Optimization of the illumina COVIDSeq^{TM} protocol for decentralized, cost-effective genomic surveillance. *Pract. Lab. Med.* 34, e00311. doi: 10.1016/j.plabm.2023.e00311

Carpenter, R. E., Tamrakar, V., Chahar, H., Vine, T., and Sharma, R. (2022). Confirming multiplex RT-qPCR use in COVID-19 with next-generation sequencing: strategies for epidemiological advantage. *Glob Health Epidemiol. Genom.* 2022, 2270965. doi: 10.1155/2022/2270965

Chen, P., Sun, W., and He, Y. (2020). Comparison of the next-generation sequencing (NGS) technology with culture methods in the diagnosis of bacterial and fungal infections. *J. Thorac. Dis.* 12 (9), 4924–4929. doi: 10.21037/jtd-20-930

Das Purkayastha, S. (2020). "Chapter 17 - diversity and the antimicrobial activity of vaginal lactobacilli: current status and future prospective," in *Recent advancements in microbial diversity*. Eds. S. De Mandal and P. Bhatt (Academic Press), 397–422. doi: 10.1016/B978-0-12-821265-3.00017-7

Das Purkayastha, S., Mrinal, K., Bhattacharya,, and Himanshu, K. (2020). "Prasad, and surajit de mandal. 2020," in *Recent advancements in microbial diversity* (Academic Press), 397–422.

Detweiler, K., Mayers, D., and Fletcher, S. G. (2015). Bacteruria and urinary tract infections in the elderly. *Urol Clin. North Am.* 42 (4), 561–568. doi: 10.1016/j.ucl.2015.07.002

Duan, W., Yang, Y., Zhao, J., Yan, T., and Tian, X. (2022). Application of metagenomic next-generation sequencing in the diagnosis and treatment of recurrent urinary tract infection in kidney transplant recipients. *Front. Public Health* 10, 901549. doi: 10.3389/fpubh.2022.901549

Finton, M. D., Meisal, R., Porcellato, D., Brandal, L. T., and Lindstedt, B. A. (2020). Whole genome sequencing and characterization of multidrug-resistant (MDR) bacterial strains isolated from a Norwegian university campus pond. *Front. Microbiol.* 11, 1273. doi: 10.3389/fmicb.2020.01273

Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N., and Weinstock, G. M. (2015). Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *mBio* 6 (6), e01888–e01815. doi: 10.1128/mBio.01888-15

Greub, G., and Raoult, D. (2002). "Actinobaculum massiliae," a new species causing chronic urinary tract infection. *J. Clin. Microbiol.* 40 (11), 3938–3941. doi: 10.1128/JCM.40.11.3938-3941.2002

Hall, D. E., and Snitzer, J. A. (1994). Staphylococcus epidermidis as a cause of urinary tract infections in children. *J. Pediatr.* 124 (3), 437–438. doi: 10.1016/s0022-3476(94)70370-1

Higgins, A., and Garg, T. (2017). An emerging cause of urinary tract infection in older adults with multimorbidity and urologic cancer. *Urol Case Rep.* 13, 24–25. doi: 10.1016/j.eucr.2017.03.022

Jhang, J. F., Hsu, Y. H., Peng, C. W., Jiang, Y. H., Ho, H. C., and Kuo, H. C. (2018). Epstein-Barr Virus as a potential etiology of persistent bladder inflammation in human interstitial Cystitis/Bladder pain syndrome. *J. Urol.* 200 (3), 590–596. doi: 10.1016/j.juro.2018.03.133

Jones-Freeman, B., Chonwerawong, M., Marcelino, V. R., Deshpande, A. V., Forster, S. C., and Starkey, M. R. (2021). The microbiome and host mucosal interactions in urinary tract diseases. *Mucosal Immunol.* 14 (4), 779–792. doi: 10.1038/s41385-020-00372-5

Kanuparthy, A., Challa, T., Meegada, S., Siddamreddy, S., and Muppidi, V. (2020). Staphylococcus warneri: skin commensal and a rare cause of urinary tract infection. *Cureus* 12 (5), e8337. doi: 10.7759/cureus.8337

Kaushik, A. M., Hsieh, K., Mach, K. E., Lewis, S., Puleo, C. M., Carroll, K. C., et al. (2021). Droplet-based single-cell measurements of 16S rRNA enable integrated bacteria identification and pheno-molecular antimicrobial susceptibility testing from clinical samples in 30 min. *Adv. Sci. (Weinh).* 8 (6), 2003419. doi: 10.1002/advs.202003419

Lee, K. W., Song, H. Y., and Kim, Y. H. (2020). The microbiome in urological diseases. *Investig. Clin. Urol.* 61 (4), 338–348. doi: 10.4111/icu.2020.61.4.338

Li, M., Yang, F., Lu, Y., and Huang, W. (2020). Identification of enterococcus faecalis in a patient with urinary-tract infection based on metagenomic next-generation sequencing: a case report. *BMC Infect. Dis.* 20 (1), 467. doi: 10.1186/s12879-020-05179-0

Martinez, R. M., Hulten, K. G., Bui, U., and Clarridge, J. E. (2014). Molecular analysis and clinical significance of lactobacillus spp. recovered from clinical specimens presumptively associated with disease. *J. Clin. Microbiol.* 52 (1), 30–36. doi: 10.1128/JCM.02072-13

McAdams, D., Wollein Waldetoft, K., Tedijanto, C., Lipsitch, M., and Brown, S. P. (2019). Resistance diagnostics as a public health tool to combat antibiotic resistance: a model-based evaluation. *PloS Biol.* 17 (5), e3000250. doi: 10.1371/journal.pbio.3000250

Mezzatesta, M. L., Gona, F., and Stefani, S. (2012). Enterobacter cloacae complex: clinical impact and emerging antibiotic resistance. *Future Microbiol.* 7 (7), 887–902. doi: 10.2217/fmb.12.61

Mouraviev, V., and McDonald, M. (2018). An implementation of next generation sequencing for prevention and diagnosis of urinary tract infection in urology. *Can. J. Urol.* 25 (3), 9349–9356.

Myint, T. M., Chong, C. H. Y., Wyld, M., Nankivell, B., Kable, K., and Wong, G. (2022). Polyoma BK virus in kidney transplant recipients: screening, monitoring, and management. *Transplantation* 106 (1), e76–e89. doi: 10.1097/TP.0000000000003801

Neugent, M. L., Hulyalkar, N. V., Nguyen, V. H., Zimmern, P. E., and De Nisco, N. J. (2020). Advances in understanding the human urinary microbiome and its potential role in urinary tract infection. *mBio* 11 (2), 218–220. doi: 10.1128/mBio.00218-20

Otto, M. (2009). Staphylococcus epidermidis–the 'accidental' pathogen. *Nat. Rev. Microbiol.* 7 (8), 555–567. doi: 10.1038/nrmicro2182

Pailhoriès, H., Tiry, C., Eveillard, M., and Kempf, M. (2018). Acinetobacter pittii isolated more frequently than acinetobacter baumannii in blood cultures: the experience of a French hospital. *J. Hosp Infect.* 99 (3), 360–363. doi: 10.1016/j.jhin.2018.03.019

Perez-Carrasco, V., Soriano-Lerma, A., Soriano, M., Gutiérrez-Fernández, J., and Garcia-Salcedo, J. A. (2021). Urinary microbiome: yin and yang of the urinary tract. *Front. Cell Infect. Microbiol.* 11, 617002. doi: 10.3389/fcimb.2021.617002

Price, T. K., Dune, T., Hilt, E. E., Thomas-White, K. J., Kliethermes, S., Brincat, C., et al. (2016). The clinical urine culture: enhanced techniques improve detection of clinically relevant microorganisms. *J. Clin. Microbiol.* 54 (5), 1216–1222. doi: 10.1128/JCM.00044-16

Price, T. K., Realegeno, S., Mirasol, R., Tsan, A., Chandrasekaran, S., Garner, O. B., et al. (2021). Validation, implementation, and clinical utility of whole genome sequence-based bacterial identification in the clinical microbiology laboratory. *J. Mol. Diagn.* 23 (11), 1468–1477. doi: 10.1016/j.jmoldx.2021.07.020

Rajabnia, M., Forghani, M. S., Hasani, S., Bahadoram, M., Mohammadi, M., and Barahman, M. (2019). Prevalence and antibiotic resistance pattern of extended spectrum beta lactamase producing escherichia coli isolated from urinary tract infection. *J. Renal Inj Prev.* 8 (2), 78–81.

Reploeg, M. D., Storch, G. A., and Clifford, D. B. (2001). Bk virus: a clinical review. *Clin. Infect. Dis.* 33 (2), 191–202. doi: 10.1086/321813

Sharma, R., Singh, P., Loughry, W. J., Lockhart, J. M., Inman, W. B., Duthie, M. S., et al. (2015). Zoonotic leprosy in the southeastern united states. *Emerg. Infect. Dis.* 21 (12), 2127–2134. doi: 10.3201/eid2112.150501

Sihra, N., Goodman, A., Zakri, R., Sahai, A., and Malde, S. (2018). Nonantibiotic prevention and management of recurrent urinary tract infection. *Nat. Rev. Urol.* 15 (12), 750–776. doi: 10.1038/s41585-018-0106-x

Smith, C. J., and Osborn, A. M. (2009). Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol. Ecol.* 67 (1), 6–20. doi: 10.1111/j.1574-6941.2008.00629.x

Sokol-Leszczynska, B., Leszczynski, P., Lachowicz, D., Rostkowska, O., Niemczyk, M., Piecha, T., et al. (2019). Corynebacterium coyleae as potential urinary tract pathogen. *Eur. J. Clin. Microbiol. Infect. Dis.* 38 (7), 1339–1342. doi: 10.1007/s10096-019-03565-4

Stapleton, A., and Stamm, W. E. (1997). Prevention of urinary tract infection. *Infect. Dis. Clin. North Am.* 11 (3), 719–733. doi: 10.1016/S0891-5520(05)70382-2

Stinnett, R. C., Kent, B., Mangifesta, M., Kadam, A., Xie, H., Stauffer, S., et al. (2021). 670. precision metagenomic (PM) sequencing outperforms conventional urine culture in detecting clinically relevant microorganisms. *Open Forum Infect. Dis.* 8 (suppl 1), 437–438. doi: 10.1093/ofid/ofab466.867

Szlachta-McGinn, A., Douglass, K. M., Chung, U. Y. R., Jackson, N. J., Nickel, J. C., and Ackerman, A. L. (2022). Molecular diagnostic methods versus conventional urine culture for diagnosis and treatment of urinary tract infection: a systematic review and meta-analysis. *Eur. Urol Open Sci.* 44, 113–124. doi: 10.1016/j.euros.2022.08.009

Wojno, K. J., Baunoch, D., Luke, N., Opel, M., Korman, H., Kelly, C., et al. (2020). Multiplex PCR based urinary tract infection (UTI) analysis compared to traditional urine culture in identifying significant pathogens in symptomatic patients. *Urology* 136, 119–126. doi: 10.1016/j.urology.2019.10.018

Zhang, L., Huang, W., Zhang, S., Li, Q., Wang, Y., Chen, T., et al. (2022). Rapid detection of bacterial pathogens and antimicrobial resistance genes in clinical urine samples with urinary tract infection by metagenomic nanopore sequencing. *Front. Microbiol.* 13, 858777. doi: 10.3389/fmicb.2022.858777

TYPE Perspective
PUBLISHED 13 July 2023
DOI 10.3389/fpubh.2023.1195581

Implementing laboratory automation for next-generation sequencing: benefits and challenges for library preparation

Jillian N. Socea†, Victoria N. Stone†, Xiaorong Qian, Paula L. Gibbs and Kara J. Levinson*

Division of Laboratory Services, Tennessee Department of Health, Nashville, TN, United States

OPEN ACCESS

EDITED BY
Peera Hemarajata,
Association of Public Health Laboratories,
United States

REVIEWED BY
Utsav Pandey,
Kaiser Permanente, United States
Kyle Rodino,
University of Pennsylvania, United States

*CORRESPONDENCE
Kara J. Levinson
✉ kara.levinson@tn.gov

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

RECEIVED 28 March 2023
ACCEPTED 15 June 2023
PUBLISHED 13 July 2023

CITATION
Socea JN, Stone VN, Qian X, Gibbs PL and Levinson KJ (2023) Implementing laboratory automation for next-generation sequencing: benefits and challenges for library preparation. *Front. Public Health* 11:1195581.
doi: 10.3389/fpubh.2023.1195581

In the wake of COVID-19, the importance of next-generation sequencing (NGS) for diagnostic testing and surveillance-based screening has never been more evident. Considering this, continued investment is critical to ensure more public health laboratories can adopt these advanced molecular technologies. However, many facilities may face potential barriers such as limited staff available to routinely prepare, test, and analyze samples, lack of expertise or experience in sequencing, difficulties in assay standardization, and an inability to handle throughput within expected turnaround times. Workflow automation provides an opportunity to overcome many of these challenges. By identifying these types of sustainable solutions, laboratories can begin to utilize more advanced molecular-based approaches for routine testing. Nevertheless, the introduction of automation, while valuable, does not come without its own challenges. This perspective article aims to highlight the benefits and difficulties of implementing laboratory automation used for sequencing. We discuss strategies for implementation, including things to consider when selecting instrumentation, how to approach validations, staff training, and troubleshooting.

KEYWORDS
next-generation sequencing, automation, validation, verification, public health

1. Introduction

The introduction of accessible next-generation sequencing (NGS) technology has changed the landscape of clinical and public health microbiology. It offers the possibility of improving diagnostics, surveillance, and public health response. Sequencing can now be used to routinely support outbreak investigations, thus helping laboratories detect disease clusters sooner and with more clarity (1–3). By replacing standard microbiology methods with culture-independent applications for pathogen detection, NGS has the potential to guide more targeted patient care (4). Therefore, it is not surprising over the last 3 years there has been a major push to invest in genomic sequencing (5). NGS data has been essential during the COVID-19 pandemic. When combined with epidemiology, it offers a means to investigate transmission patterns as the virus continues to spread across the globe. Now, more than ever, clinical, and public health facilities are working with limited staff. New hires may lack the knowledge or experience to understand sequencing assays. Thus, at first glance, implementing NGS technologies may seem too complicated and time-consuming for laboratories to onboard or to even try to increase sequencing capacity. Workflow automation provides an opportunity to overcome some of these barriers. General laboratory automation has been used in many different types of laboratories for years (6), and has only exploded more recently, resulting in a multi-billion dollar market (7). As demand for next-generation sequencing increases, it only makes sense to consider how automation could potentially be used to support this type of testing. Here we discuss several aspects of automation for preparing sequencing libraries. We highlight the key benefits as well as some of the challenges of using automated liquid handlers. We also discuss how we approached validating one of these systems.

2. The benefits of NGS automation

Preparing specimens for next-generation sequencing is a time-consuming process, involving multiple steps, starting from sample extraction. The process of preparing the sequencing libraries is a critical part in obtaining high-quality results. It involves several time-sensitive steps, pipetting small volumes, as well as washing and re-washing samples. The entire process can take hours of a bench scientist's time and one mistake can result in the loss of an entire day's worth of work. Several manufacturers have designed automated liquid handlers specifically for this complicated process. Automated instruments of various sizes and capabilities have been created and can be programmed to perform an entire library prep protocol as a single streamlined process or in separated individual steps. While automation is not a magical solution to fix every problem, it does offer several benefits worth noting and may allow laboratories to overcome some of the hurdles involved with implementing NGS (8).

2.1. Improved quality

The most obvious reason for automation is enhanced sample quality, often with greater consistency than most laboratory scientists can reproduce manually. For NGS, a lot of library prep protocols use magnetic beads and repeated wash steps for purification and fragment size selection. Manual preparation requires that scientists are skilled and fully trained, otherwise samples may be lost, contaminated, or of suboptimal quality, all which affects downstream analyses. Automated platforms are designed for these precise pipetting steps, producing consistent high-quality libraries in less time than it takes using manual preparation. In our experience, we have observed quality improvements by a few measures including more uniform nucleic acid fragment lengths and less need for repeat testing of samples. Ultimately, a decrease in failed runs saves time, reagents, and supplies.

2.2. User friendly interface

Although the backend algorithms to automate a sequencing library preparation protocol can be complicated, many platforms come with a computer pre-programmed with user-friendly control software. Scientists do not need a lot of experience with NGS or a deep understanding of the scientific process to setup or run these liquid handlers. Established protocols often use simple images to display exactly where consumables should be placed, provide visual cues to indicate what step of the process is occurring, and the instrument can perform calculations to determine reagent volumes needed for the number of samples being run. Therefore, net training time is reduced, and scientists should not need specialized programming expertise to troubleshoot basic issues.

2.3. Increased flexibility

Automated instruments often allow laboratories to scale-up or scale-down as needed. There are instruments that offer variable levels of throughput while maintaining quick turnaround times. A lab can process 4–384 samples per run, depending on their system and needed output. Another added benefit is that some platforms offer modular workflow options with safe stopping points that enable labs to adjust as needed. Instead of an end-to-end process, labs can opt to only use the instrument for certain steps like library clean-up. Those that need more than the standardized library prep protocol offered through commercial vendors, manufacturers like Agilent and Beckman Coulter have graphical or simplified software interfaces that removes some of the complexity of creating customized protocols. They also offer training courses on method programming through their software. Hamilton and Beckman Coulter also have decks that can be reconfigured for new workflows. However, this may not be the case for all platforms. Some platforms have locked-in protocols that require the manufacturer to establish new workflows.

2.4. Timesaving

Automated platforms for library prep can perform more than just liquid transfer and mixing. Instruments can be customized to include on-deck thermocyclers, shakers, and heat blocks for a fully automated system, reducing the need for any manual interference. If prepared manually, the Illumina DNA Prep protocol takes approximately 3 h to generate a sequencing ready library. While the overall run time is similar for an automated workflow, the hands-on time is far reduced (approximately 30 min to set up instrument plus 2.5 h automated run time versus 3+ hours for the manual protocol per 8 samples processed). An added benefit is that only one scientist is needed to setup an instrument, regardless of the number of samples being run. Once the samples are loaded and the program is started, that scientist is free to walk-away and focus on other tasks.

3. The challenges with NGS automation

While automated workflows have many benefits, as mentioned above, there are some significant challenges to consider before deciding to implement such systems.

3.1. System cost, design, and setup

Automated instruments are often quite expensive to purchase (quotes we have received range between $45K–300K spanning a low throughput platform and two different high throughput instruments), and careful consideration should be given to determine whether one may be realistic or necessary for the current and future workload.

Such systems have become increasingly complex, often with many add-on options like on-deck thermocyclers, bulk pipettor attachments, and robotic arms to suit different laboratory capacity needs and to perform various assay protocols. It is worthwhile to have a firm understanding of the fundamental components of a given procedure and/or product before settling on the system design that will be optimal for use. Depending on the starting sample material, reagent kit type, and sequencing platform to be used, there may be a limitation as to what instruments are compatible and available to choose from. We estimate it costs about $40 per sample, so there is likely little room to save in actual cost, but the relief in hands-on technician time may be worth it. Additionally, as automation becomes more widespread, costs may come down for consumables and for new systems in the future.

3.2. Troubleshooting and training

Initial on-site training is likely to be provided by a representative from the manufacturer to allow for familiarization with the installed instrument and to provide an overview of the basic workflow that it performs. However, hands-on experimentation will likely be required to gain experience and a better understanding of the system's nuances. Clean water runs and test runs will help assess what steps and actions may present run errors or other issues that could impact the downstream quality of testing samples. Modifications to the software program running the workflow may be necessary to ensure that steps are performed accurately and seamlessly with minimal error and stoppage within the user's laboratory. Although it's worth noting that modifications may be limited by the manufacturer. In our experience, there is very little prospect of access to off-site vendor training to enable customization of software or protocol workflows. As with any new instrument, testing personnel will need to be trained. It may prove fruitful to train more senior staff (upper management or section supervisors) as "super users" to protect against loss of expertise to employee turnover or in times of limited availability of competent staff. We recommend maintaining a minimum of two "super users" at all times. These "super users" should be proficient in troubleshooting more difficult errors that may require remote assistance from the manufacturer, as well as the ability to realign deck positions ("deck teaching"), among other skills.

3.3. Routine performance and maintenance

One of the perks of using automated workflows is the concept of being able to "walk-away" without interruption to testing. The true experience though may be more complicated than that. Following manufacturer's instructions for daily, weekly, and more long-term maintenance programs is crucially important to keep the instrument running smoothly. Routine maintenance includes channel calibration, both spacing, aspiration and dispensing, as well as surface cleaning to remove dust or other contaminates. This will likely be automatically prompted for by the instrument, but if not, a regular schedule (weekly) for such activities should be implemented. Annual preventative maintenance is also often provided by the manufacturer under special contracts (additional $15K–30K/year) to limit likelihood of bigger problems accumulating. These preventative maintenance appointments likely require scheduling with the on-site representative, which means there may be a waiting period before service is performed. The same is often true for any other service calls that may be needed when the instrument has an error or issue that the user is not able to resolve on their own. In our experience, direct communication with field engineers and applications specialists is common, which reduces instrument downtime and removes the need for tiered response through the general customer service line. Although our setup does not allow remote access, others may be able to design their system to enable this feature to limit the need for on-site visits when fixing minor errors and problems. Maintaining competency on a manual preparation method is recommended to ensure workflow is not halted if instrumentation requires repair or service.

3.4. Quality control within system limitations

As with all assays, quality controls (QC) must be continuously monitored to ensure the implemented instrument and protocol provides consistent, reliable, and accurate results. Within sequencing, there are often many QC "checkpoints" to confirm that each sample's integrity is maintained from one step to another; these are often at key points in the procedure (e.g., after DNA extraction, after library preparation, after sequencing, etc.). Within automated systems, and depending on the library preparation kit used, there may be limitations in the ability to measure the sample quality at these predetermined points. In some cases, it may be necessary to modify appropriate timepoints for such system checks, and to be creative with when and how quality can be measured. For example, certain extraction/cell lysis methods may end with a beaded product, therefore, traditional quantification methods may not be practical after such steps. If situations like this arise, it becomes critical to establish quality thresholds at the next earliest available opportunity to limit time, sample, and reagent waste. In the instance that the extracted specimen cannot be measured due to the presence of beads, we find the QC checkpoint of quantifying DNA upon completion of the library preparation to be critical in determining whether each sample meets the quality needed for sequencing. This means that there may be reagent and sample waste if one does not meet the threshold for sequencing, and the sample will have to be completely re-extracted. This can be an annoyance in our experience, but it has not happened frequently or more often than in other methods.

4. One system does not fit all

There are a variety of automated platforms currently on the market geared toward next-generation sequencing library preparation (Table 1). Before committing, laboratories should assess their budget, facilities, and sequencing workflow to help identify what would work best for them to meet their sequencing goals. While a clinical lab may prioritize high-volume testing, a research facility may require a system with a flexible workflow. For this summary, we will focus on three main areas: system compatibility, system capability, and system capacity.

TABLE 1 Fully automated library preparation platforms.

Manufacturer and platforms	
Mid- to High-Throughput (up to 384 samples)	Agilent • Bravo
	Beckman Coulter • Biomek i-Series
	Eppendorf • ep*Motion* ®
	Hamilton • Microlab STAR™ • Microlab VANTAGE™
	PerkinElmer • Sciclone G3® • Fontus™ • Zephyr®
	SPT Labtech • Firefly®
	Tecan • Fluent® • DreamPrep® • Freedom EVO®
Low-Throughput (<96)	Agilent • Magnis
	Beckman Coulter • Biomek NGeniuS
	PerkinElmer • BioQule™ NGS System
	Tecan • MagicPrep™

4.1. System compatibility

Ensuring a library preparation protocol is compatible with an automated platform can have a big impact on selection. Identifying a liquid handling system that already has the established vendor-approved workflow that matches the preparation kit that will be used can eliminate the time and energy needed to design a customized method. Illumina and New England Biolabs are two examples of companies that have partnered with leading automation manufacturers, including Agilent, Tecan, and PerkinElmer, to establish automated workflows for their library preparation kits. This opens the door for a single automated instrument to be used for several different sequencing applications. Another factor to consider if already performing NGS, is whether any internal changes were made to the manufacturer's procedure or use of a lab developed protocol. Usually, the lab end user will not be an expert in scripting automation. So, for any modifications the lab will need to discuss with the manufacturer to see if adjustments can be incorporated into the automated workflow. However, if a lab can foresee the need to continually configure or develop new protocols, it may be worth the effort to invest time in training on scripting or choose an instrument designed to support this feature. It is also important to consider what consumables can be used. Proprietary hardware and tips may be a limiting factor if items are in high demand and become backordered. The ability to use more generic plates and tips offers some flexibility.

4.2. System capabilities

Although fully integrated "walk-away" automation seems ideal because it can free up scientists for other tasks, it may not be realistic for every lab. Fully automated systems require more space, more complex algorithms, and can be costly. Partial automation can be as simple as an automated pipetting station programmed to transfer and mix reagents. This will still require more work but should cut down on the hands-on time and reduce potential errors when compared to complete manual prep. However, if looking to eliminate almost all hands-on interaction, it is best to look for all-in-one liquid handlers. As mentioned previously, these systems may include multi-channel pipetting heads, plate grippers for moving hardware across the instrument deck, orbital shakers, and plate magnets for bead clean-up steps. Additional features such as on-deck thermal cyclers or storage towers for consumables may not be standard, thus increasing costs. Some systems also incorporate an on-deck or attached sequencer, further minimizing manual interactions. However, these extra items take up deck space and may decrease sample throughput.

4.3. System capacity

Robots designed for small batch volumes may be ideal for low-volume laboratories or those that prioritize faster turnaround times. Instruments equipped to prep 96–384 samples will likely be beneficial for facilities of higher volumes that need to sequence larger batches, depending on the system's design. However, batching may result in an increase in turnaround times. Consumable use may also be a factor to consider. Automation requires a large amount of disposable hardware. And whether preparing 8 or 96 samples, some platforms may use the same amount of pipette tips, tubes, and plates. Therefore, it may be more economical to opt for sequencing in larger batches.

5. Designing regulatory-compliant validations

While NGS has become a more widely used practice, especially in the public health laboratory space, it may be useful to consider the assay design and implementation to meet regulatory compliance. Many regulatory programs (e.g., CLIA, CAP, etc.) have begun making more specific guidance (9, 10) and there are several useful resources available to strategize a successful approach (11).

The following components for validation of an automated liquid-handling instrument have been defined using a combination of best practices outlined by others, while tailoring to the instrument and DNA library preparation kit used (Table 2; Hamilton Microlab STAR and Illumina DNA Prep Kit products). It is worth noting that validations for laboratory developed tests (LDT) that are solely for surveillance purposes require a lower burden than those for diagnostic purposes to meet regulatory compliance. Consideration should be given as to how results will be used when on-boarding

TABLE 2 Example sample set and criteria for validation of WGS for major bacterial pathogens.

Validation criteria		
Accuracy	**Specimen list** • *Acinetobacter baumannii* • *Camplyobacter jejuni* • *Enterobacter cloacae* • *Escherichia coli* • *Klebsiella pneumoniae* • *Listeria monocytogenes* • *Neisseria gonorrhoeae* • *Salmonella enterica Heidelberg* • *Salmonella enterica Typhimurium* • *Salmonella enterica Arizoniae* (run control)	**Quality control metrics to meet criteria** Correct genus/species identification Coverage (20-40X depending on the genus) Identification of characteristic antimicrobial resistance (AMR) gene **Other important QC metrics** GC Content Number of Contigs 10 sample prepared and sequenced on a single run
Precision	Intra-Assay	5 samples prepared in duplicate and sequenced on single run and evaluated for QC metrics used in accuracy analysis
	Inter-Assay	5 samples prepared by 2 independent scientists, sequenced on separate MiSeq instruments, and evaluated for QC metrics used in accuracy analysis
	Method Comparison	5 samples prepared using manual and automated protocols, sequenced on single run, and evaluated for QC metrics used in accuracy analysis
Analytical sensitivity	DNA Input Range	1 ng - 10 ng*
	Limit of detection (LOD) determination	Run control subspecies ID with minimum accepted coverage (≥30X)
Analytical specificity	Secondary species abundance	≤ 1% of all sequencing reads are that of a "contaminant"

*Illumina recommends a standardized minimum DNA input (1 ng, extract concentration 0.2 ng/μL), which is required to obtain pipeline submittable sequencing read files.

sequencing tests and platforms. Additional comprehensive examples using other systems can be found to help design, develop, and implement across diverse settings and different laboratory setups (12, 13).

5.1. Accuracy

Defined here as a measure of agreement between the tested sample and a reference, assessed for the following:

- Wet lab – sequencing platform (e.g., Illumina MiSeq, Oxford Nanopore, PacBio, etc.)
- Dry lab – bioinformatics pipelines

We validated the use of one platform and compared results from two pipelines to generate the accuracy of the assay. We prepared 10 samples and sequenced them on a single run to measure accuracy of this LDT.

5.2. Precision

Defined here as a measurement of consistency between the tested sample when run multiple times, under different conditions (e.g., days, operators, sample preparations, etc.). The number of samples required to meet this criterion should be at the direction and approval of each individual laboratory's director. We utilized 5 samples to measure precision, as this was the minimum needed to test the range of organisms we test routinely, while also accounting for cost of supplies, reagents, and instrument use.

- Repeatability (Intra-assay precision) – samples tested in duplicate or triplicate within a single run

  Note: Be aware of potential sequencing biases or errors that can occur when there is too much similarity between samples.
- Reproducibility (Inter-assay precision) – samples prepared by individual operators on separate days, sequenced on the same run and/or on different runs

5.3. Sensitivity

Defined here as the limit of detection (LOD), we utilized the final concentration (High-sensitivity Qubit reading in ng/μL) of a prepared sample library that could be used to identify to species level.

5.4. Specificity

Defined here as the ability of a bioinformatics pipeline to identify contamination or interfering substances, as well as exclusion of a genus and/or species outside of those intended.

5.5. Method comparison (manual vs. automated protocols)

We added method comparison to determine if the results obtained from the new automated process differed from those using the currently validated manual preparation protocol. This was used solely to test the library preparation portion of the protocol, as the extraction method and the bioinformatics pipeline for analysis were identical between both methods.

5.6. Reference interval

Defined here as the normal value expected to correctly identify the genus and species of a given Gram-negative bacterial panel. However, this metric could be defined differently based on the desired target and intended use for result reporting. One example may be the presence or absence of a specific target gene.

5.7. Reportable range

Defined here as the output result that may be used for reporting, generally to include the genus and species identified, but may also contain serotype or other information. Depending on the use of a result, additional parameters with strict thresholds may be required, including coverage, Q30 scores, read length, etc.

6. Discussion

As we come out of the COVID-19 pandemic, it has become obvious that public health laboratories need to be ready to handle the next outbreak. The emergence of novel pathogens and the expansion of known antimicrobial resistant threats will likely balloon the test burden within public health over the coming decades. Sequencing, including automation, is just beginning to address public health needs and to aid in clinical diagnosis and treatment decisions. Working together with research, commercial, and clinical laboratories is essential to ensure a seamless transition from discovery and design to diagnosis, practice, and scaling. Advancement in NGS automation should be expected to continue, thus making new systems and instruments more prevalent, especially as they become more efficient and economical. Therefore, many public health laboratories should begin to consider the platforms and technologies that may work best for their workers, patients, and budgets.

As discussed previously, automation is essential to build testing capacity and to reduce the workload of manual test procedures. Although reliable and effective, automation can be complex and may bring new learning challenges to be of use. We recommend public health laboratories extend patience when acquiring new instrumentation, practice flexibility and generosity with time and resources that may be required for successful implementation and be communicative with others to problem-solve and troubleshoot. Automation is becoming more commonplace and there is an ever-growing network of laboratories and public health spaces that can work together to ensure the uptake and application of automation will continue to be valuable and successful.

Data availability statement

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

Author contributions

JS and VS conceptualized the topic and wrote equal sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

The authors would like to acknowledge the many talented people at the Tennessee Department of Health, Division of Laboratory Services. The authors thank Christina Moore and her team – Chanice Wilkes, Rhonda Kellem, and Austin Hargrove – for their assistance and expertise with NGS, as well as their generosity with instrumentation and reagents.

Conflict of interest

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

Publisher’s note

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

References

1. Chen X, Kang Y, Luo J, Pang K, Xu X, Wu J, et al. Next-generation sequencing reveals the progression of COVID-19. *Front Cell Infect Microbiol.* (2021) 11:632490. doi: 10.3389/fcimb.2021.632490

2. Dougherty CE, Graf E. Next generation sequencing for outbreak investigation in the clinical microbiology laboratory. *Clin Lab Sci.* (2019) 32. doi: 10.29074/ascls.119.001750

3. Malek A, McGlynn K, Taffner S, Fine L, Tesini B, Wang J, et al. Next-generation-sequencing-based hospital outbreak investigation yields insight into *Klebsiella aerogenes* population structure and determinants of carbapenem resistance and pathogenicity. *Antimicrob Agents Chemother.* (2019) 63:e02577-18. doi: 10.1128/AAC.02577-18

4. Morash M, Mitchell H, Beltran H, Elemento O, Pathak J. The role of next-generation sequencing in precision medicine: a review of outcomes in oncology. *J Pers. Med.* (2018) 8:30. doi: 10.3390/jpm8030030

5. The White House. (2021). Available at: https://www.whitehouse.gov/briefing-room/legislation/2021/01/20/president-biden-announces-american-rescue-plan/ (Accessed March 22, 2023).

6. Antonios K, Croxatto A, Culbreath K. Current state of laboratory automation in clinical microbiology laboratory. *Clin Chem.* (2022) 68:99–114. doi: 10.1093/clinchem/hvab242

7. Laboratory Automation Systems Global Market Report. The business research company. (2023). Available at: https://www.reportlinker.com/p06280847/Laboratory-Automation-Systems-Global-Market-Report.html (Accessed March 22, 2023).

8. Singh RR, Luthra R, Routbort MJ, Patel KP, Medeiros LJ. Implementation of next generation sequencing in clinical molecular diagnostic laboratories: advantages, challenges and potential. *Expert Rev Precis Med Drug Dev*. (2016) 1:109–20. doi: 10.1080/23808993.2015.1120401

9. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med*. (2015) 139:481–93. doi: 10.5858/arpa.2014-0250-CP

10. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J Clin Microbiol*. (2016) 54:2857–65. doi: 10.1128/JCM.00949-16

11. Centers for Disease Control and Prevention. The next generation sequencing quality initiative. (2022). Available at: https://www.cdc.gov/labquality/ngs-quality-initiative.html (Accessed March 22, 2023).

12. Kozyreva VK, Truong C-L, Greninger AL, Crandall J, Mukhopadhyay R, Chaturvedi V. Validation and implementation of clinical laboratory improvements act-compliant whole-genome sequencing in the public health microbiology laboratory. *J Clin Microbiol*. (2017) 55:2502–20. doi: 10.1128/JCM.00361-17

13. Lepuschitz S, Weinmaier T, Mrazek K, Beisken S, Weinberger J, Posch AE. Analytical performance validation of next-generation sequencing based clinical microbiology assays using a K-mer analysis workflow. *Front Microbiol*. (2020) 11:1883. doi: 10.3389/fmicb.2020.01883

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 13 July 2023
DOI 10.3389/fpubh.2023.1198189

OPEN ACCESS

EDITED BY
Shangxin Yang,
University of California, Los Angeles,
United States

REVIEWED BY
Annabelle De St. Maurice,
Los Angeles County Department of Public
Health, United States
Kin-Ming (Clement) Tsui,
University of British Columbia, Canada
Erin Young,
Utah Department of Health and Human
Services, United States

*CORRESPONDENCE
Mark Pandori
✉ mpandori@med.unr.edu
David Hess
✉ dhess@unr.edu

RECEIVED 31 March 2023
ACCEPTED 06 June 2023
PUBLISHED 13 July 2023

CITATION
Gorzalski A, Ambrosio FJ III, Massic L, Scribner MR, Siao DD, Hua C, Dykema P, Schneider E, Njoku C, Libuit K, Sevinsky JR, Van Hooser S, Pandori M and Hess D (2023) The use of whole-genome sequencing and development of bioinformatics to monitor overlapping outbreaks of *Candida auris* in southern Nevada.
Front. Public Health 11:1198189.
doi: 10.3389/fpubh.2023.1198189

The use of whole-genome sequencing and development of bioinformatics to monitor overlapping outbreaks of *Candida auris* in southern Nevada

Andrew Gorzalski[1], Frank J. Ambrosio III[2], Lauryn Massic[1], Michelle R. Scribner[2], Danielle Denise Siao[1], Chi Hua[3], Phillip Dykema[3], Emily Schneider[3], Chidinma Njoku[4], Kevin Libuit[2], Joel R. Sevinsky[2], Stephanie Van Hooser[1], Mark Pandori[1,5,6]* and David Hess[1,5]*

[1]Nevada State Public Health Laboratory, Reno, NV, United States, [2]Theiagen Consulting LLC, Highlands Ranch, CO, United States, [3]Division of Disease Control and Health Statistics, Washington State Department of Health, Public Health Laboratories, Shoreline, WA, United States, [4]Nevada Department of Health and Human Services, Las Vegas, NV, United States, [5]Department of Pathology and Laboratory Medicine, University of Nevada, Reno School of Medicine, Reno, NV, United States, [6]Department of Microbiology and Immunology, University of Nevada, Reno School of Medicine, Reno, NV, United States

A *Candida auris* outbreak has been ongoing in Southern Nevada since August 2021. In this manuscript we describe the sequencing of over 200 *C. auris* isolates from patients at several facilities. Genetically distinct subgroups of *C. auris* were detected from Clade I (3 distinct lineages) and III (1 lineage). Open-source bioinformatic tools were developed and implemented to aid in the epidemiological investigation. The work herein compares three methods for *C. auris* whole genome analysis: Nullarbor, MycoSNP and a new pipeline TheiaEuk. We also describe a novel analysis method focused on elucidating phylogenetic linkages between isolates within an ongoing outbreak. Moreover, this study places the ongoing outbreaks in a global context utilizing existing sequences provided worldwide. Lastly, we describe how the generated results were communicated to the epidemiologists and infection control to generate public health interventions.

KEYWORDS

***Candida auris*, epidemiology, whole-genome sequencing, bioinformatics, emerging pathogens**

1. Introduction

Candida auris was first identified in 2009 in Japan and has quickly become an emerging global pathogen (1). Since its discovery it has rapidly spread worldwide (2–4). Genomic analysis of *C. auris* has identified four main clades in addition to a rarer fifth clade (5). *C. auris* clades were initially identified by geographical region—Clade 1 (South Asian), Clade 2 (East Asian), Clade 3 (African), Clade 4 (South American) and Clade 5 (Middle Eastern) (2, 5). However, outside of Clade 5, all other clades have escaped their initial geographic boundaries (6).

C. auris presents multiple medical and public health challenges which contribute to its concern as an emerging pathogen. Firstly, *C. auris* commonly possesses resistance to existing antifungal pharmaceuticals (2). Secondly, *C. auris* has the ability to colonize hosts both internally and externally and often asymptomatically. This facilitates spread (6, 7) and obfuscates screening strategies. Thirdly, these traits facilitate establishment and spread within health care facilities which has prompted agencies around the world (Centers for Disease Control [CDC], European Centre for Disease Prevention and Control [EDPC] and Public Health England) to release clinical alerts on *C. auris* (8–10). Fourthly, crude estimates of mortality for hospitalized patients with candidemia is 30–72% in frequently hospitalized indivivduals.

Because of the ability of *C. auris* to spread and colonize in health care facilities, rapid identification and genomic analysis are necessities in containing outbreaks. In this study we applied robust genomic sequencing analysis to a major outbreak of *C. auris* in southern Nevada. Analysis revealed two genomically distinct, simultaneous *C. auris* outbreaks that initiated with chronological proximity. Whole genome sequencing was performed on 208 isolates associated with the outbreak. The sequences generated were utilized both to develop and to assess novel tools. These tools were utilized for identification and for phylogenetic analysis to aid the epidemiologic investigation. Three existing methodologies for analyzing *C. auris* whole genome sequences were studied and the results are shown: Nullarbor (12, 13), mycosnp (3) and TheiaEuk (14). All methods showed the ability to identify *C. auris* based on whole genome sequencing and to generate relatedness metrics. Using these tools, we describe the development of a custom, shared single-nucleotide polymorphism (SNP) method that may provide significant aid in the use of *C. auris* genomic sequences in epidemiologic investigations.

2. Materials and methods

2.1. Collection of specimens

Specimens were isolated from clinical samples collected in Nevada from August 2021 to July 2022. Two additional isolates of interest from Nevada are included in this study from 2022-11-18 and 2023-01-15. *C. auris* is not a reportable organism in Nevada, so initial clinical samples were obtained in collaboration with our ARLN lab in Washington State and southern Nevada clinical partners. Since, *C. auris* is not a reportable organism, so it is difficult to estimate the number of cases compared to the number sequenced in this timeframe. However, this study sequenced every *C. auris* isolate from the time range noted that the Nevada State Public Health Laboratory was able to obtain a cultured isolate. All relevant information for each *C. auris* isolates including clade designation, Sequence Read Archive identifier at NCBI, antifungal MICs, etc. is included in Supplementary Table S1.

2.2. Whole-genome sequencing of *Candida auris*

Genomic DNA used for sequencing was extracted using a combination of bead-beating (FastPrep-24, MP Biomedicals, Irvine, CA) and magnetic-bead purification (Maxwell RSC 48, Promega, Madison, WI). First, isolates from Sabouraud Dextrose agar plates were mixed with silica beads (Lysing Matrix C, MP Biomedical) and then mechanically sheared with 2 cycles at 6.0 m/s for 30 s with a 5 min pause between (FastPrep-24, MP Biomedical). Genomic DNA was extracted using PureFood Pathogen Kit (Promega) on a Maxwell RSC 48 (Promega) using manufacturer's protocol. Genomic DNA was library prepped using DNA Prep Kit (Illumina, San Diego, CA) using manufacturer's recommended protocol using a STARlet automated liquid handler (Hamilton Company, Reno, NV). Paired-end sequencing (2×151) was performed using Illumina's MiniSeq and NovaSeq 6,000 to a minimum depth of 35x average coverage.

2.3. Antibiotic susceptibility testing

Candida auris AST was performed using microbroth dilution and predefined gradient of antibiotic concentrations (Etest) methods. A patient isolate was grown on Sabouraud Dextrose agar plate and incubated at 30°C in ambient air for 24 h and used to make 0.5 McFarland inoculum suspension in demineralized sterile water. The 0.5 McFarland suspension was measured by spectrophotometer to verify the 0.5 McFarland (80–82% transmittance). Twenty microliters of 0.5 McFarland suspension were added into 11 mL of RPMI broth tube, and 100 μL of the RPMI diluted sample was distributed to each well of a 96-well plate pre-loaded with antibiotics and incubated along with control plates for 24 h at 35°C. The same 0.5 McFarland inoculum suspension was used to inoculate a RPMI agar plate using a sterile cotton swab. A single Amphotericin B Etest strip was applied to middle of the agar surface using sterile forceps and incubated along with control plates for 24 h at 35°C. The AST of the microbroth dilution panel was read using parabolic magnifying mirror to determine the MIC (lowest concentration where there is ≤50% growth compared to growth control well). For the Amphotericin B Etest, MIC was interpreted at value where there is 100% growth inhibition (number above where ellipse intercepts Etest strip).

2.4. Nullarbor implementation

Nullarbor is a reads-to-report bioinformatics pipeline originally written in Perl. In the Terra workflow version, Nullarbor is implemented as a single task using the Workflow Description Language (WDL). Reads are accepted in two separate arrays for read file one and read file two (n= 16). A tsv input file is generated by iterating through the arrays of read files, and this sample sheet tsv is ultimately passed into the Nullarbor analysis module. Additional inputs include an array of sample names, and a reference genome. The clade specific reference genome should be used, meaning clade must be discerned prior to running this workflow, as there is no clade typing module.

Read cleaning is performed removing sequencing adaptors and low-quality input sequencing data using Trimmomatic (15). Species identification is performed using Kraken 2 with the EuPathDB64 database available here[1] (16). *De novo* assembly is performed using SKESA (17). In addition, sequencing reads are aligned to a user-provided reference genome using Snippy, and the core phylogeny and SNP matrix are produced using snippy-core (18).

[1] https://benlangmead.github.io/aws-indexes/k2

2.5. MycoSNP implementation

MycoSNP is an open-source bioinformatics pipeline designed to call variants and construct a phylogeny from mycotic pathogen next generation sequencing data (3). The original version of this tool was written in Nextflow and implemented by the CDC Mycotics Disease Branch[2] (19). The components of this tool are wrapped in docker containers. Each of these components is an established bioinformatics method, and output files are in standard format so as to allow compatibility with downstream analytical modules. The inputs to this workflow include the raw read FASTQ files from an Illumina paired end sequencing run and a reference genome in FASTA file format. The reference genomes utilized were the CDC clade 1 reference genome [strain B11205] (GenBank Accession GCA_016772135.1) and the CDC clade 3 reference genome [strain B11221] (GenBank Accession GCA_002775015.1) (5).

MycoSNP was run with default settings as described by Bagal et al (3). The first step of the pipeline is to prepare the reference genome for alignment by masking repeat regions using nucmer[3] and generating an index for efficient alignment with the Burrows-Wheeler Aligner (BWA).[4] Next, the FASTQ files are processed and checked for quality. For FASTQ processing, SeqKit[5] is used to filter unpaired reads, SeqTK[6] is used to downsample reads, and FaQCs[7] is used to perform quality checks and read trimming. After processing, the reads are aligned to the reference genome using BWA. The resulting binary alignment map (BAM) files are sorted with SAMTools[8] and processed to remove duplicates, ensure mate-paired read information is correct, and add read groups with Picard.[9] This final step of the alignment process is to perform additional quality checks using FastQC[10] and MultiQC.[11] Variants are called using GATK.[12] The resulting GVCF files from each sample are then combined into a single VCF file, which is then filtered based on normalized variant quality, Phred-scaled probability of strand bias, mapping quality of all reads at the variant site, and the number of filtered reads that support each of the alleles found at the variant site.[13] The combined and filtered VCF is then split into individual sample-specific VCF files. Using BCFTools[14] and SeqTK, a consensus sequence is generated for each sample, and these sequences are combined into a multi-FASTA to be used as the input to the phylogenetic tree construction tools.

Multiple phylogenies are generated in MycoSNP. The phylogenetic inference tools rapidNJ,[15] FastTree2,[16] RaxML,[17] and IQTree[18] are all utilized in this final step of the workflow.

To make this workflow available on the Terra platform, the original pipeline has been split into two separate tools, each wrapped in a WDL workflow. The two new workflows perform variant calling and phylogenetic analysis independently, but the underlying components are the same as the original MycoSNP.

2.6. TheiaEuk implementation

The TheiaEuk_PE workflow performs the assembly, quality assessment, and genomic characterization of fungal genomes (14). This cloud-native workflow is implemented in the Workflow Description Language and has been operationalized on the Terra.bio platform. TheiaEuk_PE has been fashioned to accept Illumina paired-end sequencing data as the primary input but offers many optional inputs to allow the user to specify parameters for all internal components of the workflow. Input reads are preprocessed with a raw-read quality assessment followed by read cleaning (quality trimming and adapter removal), and then an additional quality assessment of the cleaned reads. Subsequently, *de novo* assembly is performed using the Shovill package with SKESA set as the default assembler. SKESA is implemented using default parameters. Once the assembly has been generated an assembly quality assessment is performed using QUAST. Using the assembly, species taxon identification is performed by GAMBIT (20). The GAMBIT implementation in TheiaEuk_PE uses a custom fungal database containing 5,667 genomes and 245 species. For some taxa identified, taxa-specific sub-workflows will be automatically activated, launching additional taxa-specific characterization tools, including a GAMBIT-based clade-typing tool and antifungal resistance detection performed using Snippy variant calling with a custom query for genes in which there are known antifungal-resistance conferring mutations (14). For *C. auris*, TheiaEuk queries the Snippy results for strings matching the *FKS1*, *ERG11*, and *FUR1* genes.

2.7. Benchmarking against other workflows

Three workflows were compared in this study using two sets of *C. auris* reads. The first set was 60 samples from clade 1 and the second set was 148 samples from clade 3. TheiaEuk was combined with kSNP3 to produce phylogenetic trees and SNP matrices. First, TheiaEuk was used to produce assemblies which were then used as inputs to the kSNP3 workflow to produce a pair of phylogenetic trees and SNP matrices. MycoSNP_Variants was used to produce VCF files which were fed into the MycoSNP_Tree workflow to produce a set of phylogenetic trees and a SNP matrix. Nullarbor was run as a single workflow producing a SNP matrix and a phylogenetic tree. Each VM deployed to run these workflows was

[2] https://github.com/CDCgov/mycosnp-nf
[3] https://github.com/garviz/MUMmer/blob/master/nucmer
[4] https://github.com/lh3/bwa
[5] https://github.com/shenwei356/seqkit
[6] https://github.com/lh3/seqtk
[7] https://github.com/LANL-Bioinformatics/FaQCs
[8] https://github.com/samtools/
[9] https://github.com/broadinstitute/picard
[10] https://github.com/s-andrews/FastQC
[11] https://github.com/ewels/MultiQC
[12] https://github.com/broadinstitute/gatk
[13] https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants
[14] https://github.com/samtools/bcftools/releases
[15] https://github.com/johnlees/rapidnj
[16] https://github.com/citiususc/veryfasttree
[17] https://github.com/stamatak/standard-RAxML
[18] https://github.com/Cibiv/IQ-TREE

given runtime parameters of 32 cpus and 128 GB of memory. These compute resources were allocated to each VM, so workflows that launched several VMs simultaneously took advantage of parallelization.

2.8. *Candida auris* specific subroutines within TheiaEuk

Upon the taxonomic assignment of *C. auris* to a sample, TheiaEuk_PE automatically triggers two taxa-specific sub-workflows (14). First, a clade-typing workflow is launched. Clade-typing is performed using a modified version of the GAMBIT module to determine which of the five clade specific references most closely matches the query sequence. The output of the clade-typing module includes the clade assignment as well as a clade-specific annotated genome which is then passed to the antifungal resistance detection module. Snippy is used to align reads to the annotated reference genome and call variants. The variants are annotated with the genes in which they are found because the input reference genome is annotated. The variants are then queried for any that occur within genes known to contain resistance conferring mutations. This method is used rather than reporting only known resistance conferring mutations to ensure that novel resistance conferring mutations are not ignored.

2.9. Shared SNP analysis

This analysis uses the VCF file from kSNP3 which lists each unique SNP in a dataset with the 9 base pairs upstream and downstream of the SNP location (21). The SNP output for clade 1 is made against the CDC clade 1 reference genome [strain B11205] (GenBank Accession GCA_016772135.1) and the SNP output for clade 3 is made against the CDC clade 3 reference genome [strain B11221] (GenBank Accession GCA_002775015.1) (5). The VCF file displays if each SNP is present, absent or unassembled for each input genomes. This analysis focuses only on SNPs that are assembled in each input genome and then filters out from that group SNPs that are in every input genome (save the reference genome) and SNPs that are unique to only one of the input genomes. The SNPs that remain are referred to as "shared SNPs," falling somewhere between unique and present in all genomes in your query set. These SNPs are manually clustered to form groups that have unique patterns of shared SNPs. These SNPs were not annotated in our analysis.

3. Results

3.1. NV outbreak

An outbreak of *C. auris* in southern Nevada was first detected in August 2021. As of October 31st, 2022, over 500 cases had been reported with over 200 isolates preserved. We report on 210 isolates including 2 isolates that represent new introductions to Nevada as detected by our analysis pipelines described below. All isolates were subjected to whole genome sequencing.

3.2. Global relatedness of southern Nevada clades to global clades

To initially assess isolates associated with the outbreak genomically, we utilized a phylogenetic tree-based comparison on the entire genome against a subsampling of previously submitted clade 1 or clade 3 strains from organizations around the world (Figures 1A,B; Supplementary Table S2) (21). These trees establish that the southern Nevada strains have a unique phylogenetic signature among all *C. auris* isolates previously submitted to public repositories. Figure 1A presents *C. auris* clade 1 samples that have been sequenced and uploaded to public repositories. The major phylogenetic groups are highlighted with different colors and annotated by the region where the *C. auris* isolates were collected with the Nevada isolates highlighted in purple. The Nevada clade 1 outbreak is genetically distinct from other outbreaks in the U.S. as shown in Figure 1A. The index southern Nevada case for the clade 1 outbreak was SRR23249008 (Supplementary Table S1).

A similar analysis was performed with clade 3 samples and is shown in Figure 1B. As with clade 1 analysis, it became clear that the Nevada clade 3 samples (highlighted in purple) were genetically distinct from other previously sequenced outbreaks. The index southern Nevada cases for the clade 3 outbreak were SRR19738700 and SRR23109087 which were both collected on 11-02-2021 in the case facility (Supplementary Table S1). Note the one isolate labeled Arizona01 in the shaded purple box was collected in Arizona from a southern Nevadan patient. Epidemiological investigation strongly suggested the patient contracted *C. auris* in southern Nevada prior to travel to Arizona (data not shown). The information on the case described in the previous sentence was obtained and shared with the Nevada State Public Health Laboratory in collaboration with our public health partners in Utah and Nevada. These data were collected and shared in accordance with IRB protocols. We concluded having a pipeline(s) of rigorous bioinformatic tools capable of handling fungal microbes would be necessary for the public health response to these simultaneous and distinct outbreaks occurring in southern Nevada.

3.3. State of fungal bioinformatic whole-genome sequencing pipelines *circa* march 2022

After an initial analysis of the outbreak specimens with regard to clade, we sought to further determine the utility of phylogenetic analysis to assist disease control. Upon initiation of sequencing and genomic analysis of the outbreak at the Nevada State Public Health Laboratory, one computational method /pipeline was available for assessments of *C. auris* (12, 13). Another was completed but unpublished (3). Difficulties with implementation of the former led our group to develop a novel pipeline for identification and phylogenetic analysis of *C. auris* genomes. This pipeline has been named "TheiaEuk" (14). As we sought to determine the best methods for using sequencing to assist disease control efforts for this outbreak, we sought to compare all three methods in terms of capability and functionality.

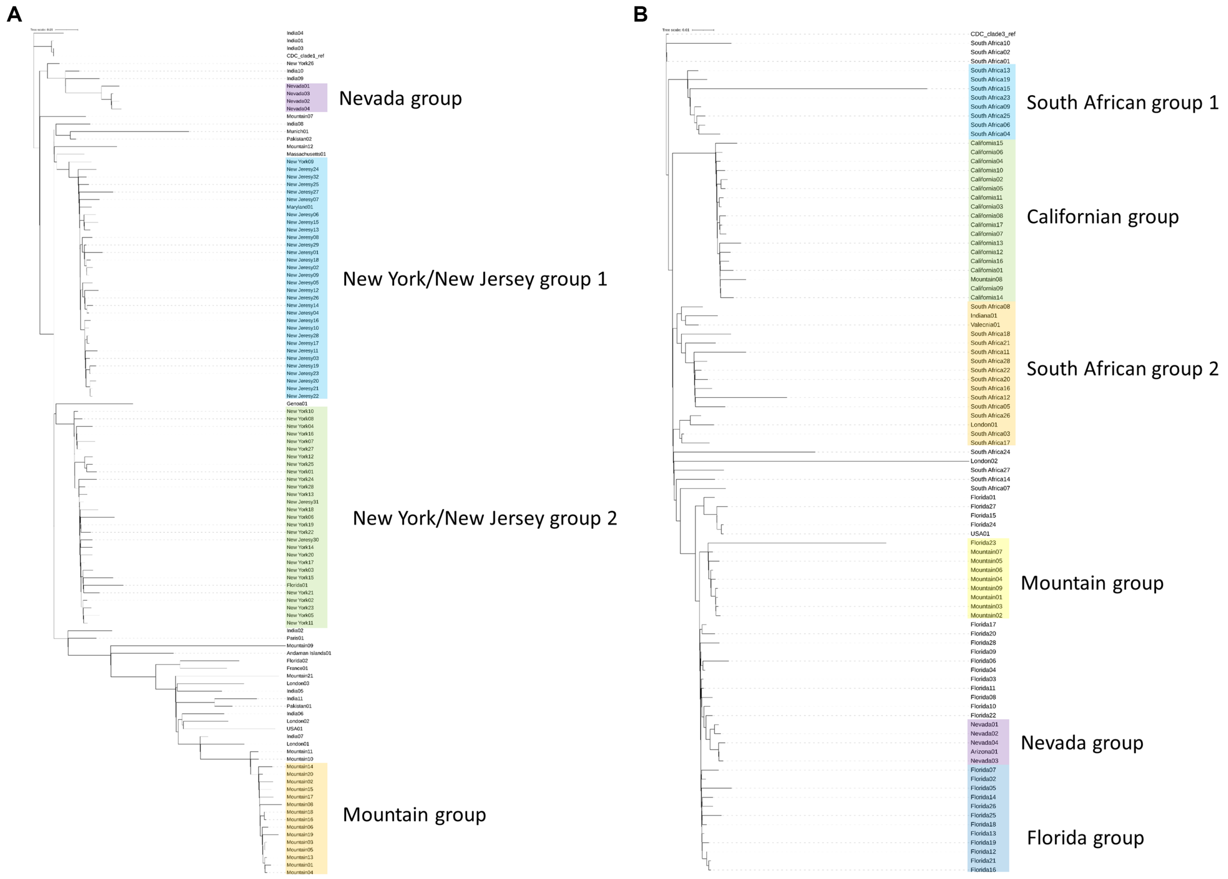


FIGURE 1
Phylogenetic trees generated by kSNP3 (21) on *Candida auris* isolates. Isolates are labeled by region from which they were isolated. Samples for each tree are listed with their SRA ID in Supplemental Table 2. Shading on the tree represents highly related branches which clustered by geography. **(A)** Phylogenetic tree of Clade 1 isolates including 4 isolates from the southern Nevada outbreak (colored in purple). **(B)** Phylogenetic tree of Clade 3 isolates including 5 isolates from the southern Nevada outbreak (colored in purple).

3.4. Pipeline comparisons for *Candida auris* genome assemblies

In testing three methods, we assembled genomes ($n=60$) from clade I from patients found infected by *C. auris* in southern Nevada. Assembly and downstream analyses were completed using each of three workflows: Nullarbor, TheiaEuk and Mycosnp. Upon completion, results from Nullarbor needed no additional analysis. TheiaEuk and Mycosnp required additional step for the generation of SNP matrices and/or phylogenetic trees (Figure 2). Times required for analyses are shown in Table 1. As shown in Table 2, all methods assemble genomes of nearly identical sizes with the average genome lengths of Nullarbor being 12,276,509 bp, TheiaEuk being 12,288,829 bp and MycoSnp being 12,406,106 bp. For assembly, Nullarbor uses Shovill v1.1.0 with SKESA v2.4.0 as the default setting. TheiaEuk uses Shovill v1.1.0 with SKESA v2.4.0 as the default setting. MycoSNP uses a reference guided assembly that produces the same genome length for each sample using method BWA v0.7.17 for read alignment and GATK v4.2.5.0 for variant calling. MycoSNP using reference guided assembly creates a single contig per chromosome where Nullarbor produces an average of 683 contigs from our Clade I samples and TheiaEuk produces an average of 505 contigs from tested clade I samples (Table 2).

3.5. Benchmarking TheiaEuk, MycoSNP and Nullarbor

Comparison of the three whole-genome sequencing pipelines based on analysis time was performed on the same test set described in the previous section (Table 1). All pipelines were run with the same virtual machines (Materials and Methods). MycoSNP had the fastest report time at 2 h and 5 min, followed by TheiaEuk at 8 h and 10 min with nullarbor requiring 26 h and 12 min.

3.6. Pipeline comparisons for *Candida auris* SNP matrices

Each method produces SNP matrices which display the calculated number of SNPs between each sample in an analyzed set. We compared the number of SNPs detected by each method compared to the first clade I sample by numerical order based on our internal nomenclature (SRR19664611) to all other samples. We then calculated the absolute differences between each pairwise SNP comparison between two methods. Comparing Nullarbor and TheiaEuk the difference was 1.9 (± 2.1) SNPs with Nullarbor consistently reporting fewer SNPs. Comparing Nullarbor and MycoSNP the difference was 1.9 (± 2.2)

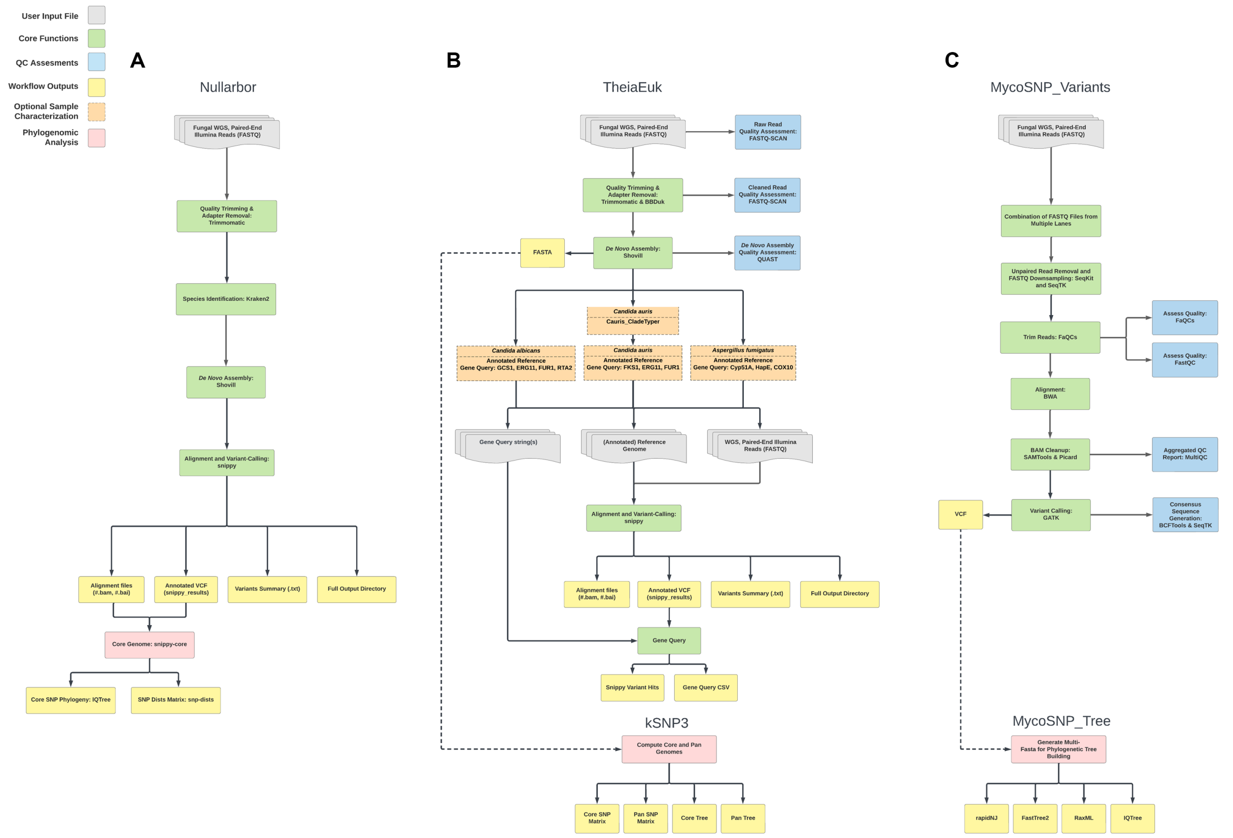

User Input File
Core Functions
QC Assesments
Workflow Outputs
Optional Sample Characterization
Phylogenomic Analysis
A
Nullarbor
Fungal WGS, Paired-End Illumina Reads (FASTQ)
Quality Trimming & Adapter Removal: Trimmomatic
Species Identification: Kraken2
De Novo Assembly: Shovill
Alignment and Variant-Calling: snippy
Alignment files (#.bam, #.bai)
Annotated VCF (snippy_results)
Variants Summary (.txt)
Full Output Directory
Core Genome: snippy-core
Core SNP Phylogeny: IQTree
SNP Dists Matrix: snp-dists
B
TheiaEuk
Fungal WGS, Paired-End Illumina Reads (FASTQ)
Raw Read Quality Assessment: FASTQ-SCAN
Quality Trimming & Adapter Removal: Trimmomatic & BBDuk
Cleaned Read Quality Assessment: FASTQ-SCAN
FASTA
De Novo Assembly: Shovill
De Novo Assembly Quality Assessment: QUAST
Candida auris
Cauris_CladeTyper
Candida albicans
Annotated Reference
Gene Query: GCS1, ERG11, FUR1, RTA2
Candida auris
Annotated Reference
Gene Query: FKS1, ERG11, FUR1
Aspergillus fumigatus
Annotated Reference
Gene Query: Cyp51A, HapE, COX10
Gene Query string(s)
(Annotated) Reference Genome
WGS, Paired-End Illumina Reads (FASTQ)
Alignment and Variant-Calling: snippy
Alignment files (#.bam, #.bai)
Annotated VCF (snippy_results)
Variants Summary (.txt)
Full Output Directory
Gene Query
Snippy Variant Hits
Gene Query CSV
kSNP3
Compute Core and Pan Genomes
Core SNP Matrix
Pan SNP Matrix
Core Tree
Pan Tree
C
MycoSNP_Variants
Fungal WGS, Paired-End Illumina Reads (FASTQ)
Combination of FASTQ Files from Multiple Lanes
Unpaired Read Removal and FASTQ Downsampling: SeqKit and SeqTK
Trim Reads: FaQCs
Assess Quality: FaQCs
Assess Quality: FastQC
Alignment: BWA
BAM Cleanup: SAMTools & Picard
Aggregated QC Report: MultiQC
VCF
Variant Calling: GATK
Consensus Sequence Generation: BCFTools & SeqTK
MycoSNP_Tree
Generate Multi-Fasta for Phylogenetic Tree Building
rapidNJ
FastTree2
RaxML
IQTree

FIGURE 2
Workflow comparisons of whole-genome sequencing bioinformatic pipelines that analyze *C. auris* WGS data. The shaded key highlights the major steps performed in analysis for ease of comparison. For details on each workflow see Materials and Methods. **(A)** Nullarbor **(B)** TheiaEuk **(C)** MycoSNP_variants.

TABLE 1 Run times for each of the tested WGS pipelines on both the Clade 1 and Clade 3 isolate sets.

	Clade 1 [time (hr:min)]	Clade3 [time (hr:min)]
TheiaEuk (Average)	2:25	2:47
kSNP3	1:09	3:03
Total	3:34	5:50
MycoSNP (Average)	7:09	8:19
MycoSNP_Tree	0:24	1:11
Total	7:33	9:30
Nullarbor	25:16:00	53:44:00

For Clade 1 $n = 60$ and for Clade 3 $n = 148$.

TABLE 2 For the Clade 1 dataset (n=60) the comparison of WGS assembly stats (genome length and number of contigs) is reported.

	Mean genome size (bps)	Number of contigs
Nullarbor	12,276,509 (± 27,775)	683 (± 215)
TheiaEuk	12,288,829 (± 25,062)	505 (± 248)
MycoSNP	12,406,106	NA

SNPs with Nullarbor consistently reported fewer SNPs. Comparison of TheiaEuk and MycoSNP resulted in a difference of 0.57 (± 0.89) SNPs with MycoSNP consistently reporting more SNPs (Table 3).

3.7. Development of a pipeline to distinguish fine grain differences in ongoing outbreaks

Distinguishing genetically related isolates within an outbreak can be challenging for pathogens with low rates of mutation (22). A SNP can be the result not only of biological introduction, but also introduced through sequencing and biocomputational methods. Within the outbreak observed herein, core genome assemblies possess a large number of shared SNPs when compared to the CDC clade references and have relatively smaller number of distinguishing mutations that define subgroups (Figure 3A). For example, all but a single clade 1 isolate share 52 common SNPs (Supplemental data), yet Table 4 shows that the most common clade 1 subclade (Group K) differs by only 3 SNPs from the second most common subgroup (Group B). We propose that the usage of subsets of *shared* mutations that follow asexual microbial evolution theory would define the most highly related subgroups (Figure 3A). To this end it can be observed that within a clade, most isolates which share a large number of "core" SNPs compared to the CDC reference, show relatedness relevant to epidemiological investigation. Such cases, then with additional SNPs shared, result in cases with a distinct profile of descendancy and thus would be assumed to have the highest level of relatedness (Table 4).

TABLE 3 Calculated SNP differences between the WGS pipelines for Clade 1 (*n*=60).

Assembly Method #1	Assembly Method #2	Ave Difference in SNP Matrix (± SNPs)
MycoSNP	TheiaEuk	0.57 (± 0.89)
MycoSNP	Nullarbor	1.9 (± 2.2)
TheiaEuk	Nullarbor	1.9 (± 2.1)

3.8. Inference of relatedness among *Candida auris* clade I outbreak samples

In order to utilize whole-genome variation to provide disease control investigators with data regarding the most related sets of isolates, specific SNPs were studied for evidence of inheritance patterns. Clade I ($n = 60$) isolates were analyzed using kSNP3 using the clade I CDC reference strain [strain B11205] as our reference (21). This analysis generated 208 SNPs whereupon the parent (defined as the reference strain) or variant sequence was detected in all 60 strains. Of these SNPs, 109 were present in only one isolate (aka unique SNPs). Of the 99 SNPs that were shared among two or more Clade I isolates, 54 were present in 57 out of 60 clade I isolates. The remainder of the analysis focused upon these 57 strains which all shared this set of "core NV clade I" SNPs.

Among these 57 strains 19 SNP sets were identified which differentiated this group based on genetics. These SNP sets (Table 4) contained between 1 and 8 SNPs. The presence of the SNP set in an isolate is represented with a "+" in Table Y and its absence is represented with a "-". We identified 15 groups of isolates that had more than one member and had a unique combination of SNP sets (Table 4)—we designated these as clade 1 Groups A through O. Provision of group designations was to assist disease controllers in Nevada in having a nomenclature to describe cases. It was not an attempt to create a novel general nomenclature field-wide. Lastly, we present a subset of our inferred clade 1 transmission network based on these data in Figure 3B.

3.9. Inference of relatedness among *Candida auris* clade III outbreak samples

The analysis described above was repeated for the 148 Clade III isolates using the clade III CDC reference strain [strain B11221] as our ancestral outgroup for kSNP3. This analysis generated 401 SNPs where the parent or variant sequence was detected in all 148 strains. Of these SNPs, 280 were present in only one isolate (aka unique SNPs). Of the 121 SNPs that were shared among two or more clade III isolates, 28 were present in 147 out of 148 clade III isolates. Focus was placed upon these 147 strains which all shared this set of "core NV clade III" SNPs.

Among these 147 isolates 42 SNP sets were identified which differentiated this group based on genetics. These SNP sets (Supplementary Table S3) contained between 1 and 5 SNPs. The presence of the SNP set in an isolate is represented with a "+" in Supplementary Table S3 and its absence is represented with a "-". Groups of isolates ($n = 35$) were identified that had more than one member and had a unique combination of SNP sets (Supplementary Table S3)—these were designated as Clade 3 Groups A through KK.

3.10. Discovery of new introductions to *Candida auris* to southern Nevada using shared SNP analysis

The shared SNP analysis described above is performed by the Nevada State Public Health Laboratory on a regular basis since September 2022. During that time, this analysis has identified two novel clade 1 introductions. These novel introductions have a unique set of "core" SNPs that are different from the Southern Nevada "core" SNP signature. To quantify this, we ran kSNP3 with four members of the southern Nevada Clade 1 outbreak with the suspected two novel clade 1 introductions (21). The first novel introduction represented by isolate SRR23137821 has 2,519 SNPs not shared by any of the original clade 1 isolates from this outbreak (Table 5). The second novel introduction represented by isolate SRR23920687 has 87 SNPs not shared by any of the original clade 1 isolates from this outbreak (Table 5). While all three have a small, overlapping set of 10 common SNPs when compared to the CDC clade 1 reference strain, the vast genetic diversity detected by the shared SNP analysis shows that these are new introductions.

4. Discussion

Candida auris is among the most challenging of healthcare-associated infections (2, 6–11). It combines the ability to persist environmentally with inherent drug resistance and the ability to cause significant morbidity and mortality. As such, public health must bring every tool at its disposal to bear on this threat. Herein, we assess one possible tool for its ability to assist in combatting *C. auris* outbreaks: genomics, in response to multiple, complex outbreaks in Nevada, we sought to generate as much genomic intelligence as possible to better understand the spread of the pathogen. The use of genomics to track and to describe pathogens is certainly not novel. However, its application to *C. auris* outbreaks is relatively new. As a fungal pathogen, *C. auris* has been shown to have a mutation rate much slower than other, healthcare associated agents (23–26). Slower mutation rates may result in less diversity in outbreak populations, thus limiting the ability to distinguish cases within transmission networks. Confronting this, we assessed the genomic diversity of whole genome sequences from numerous isolates associated with outbreaks of clade I and clade III. The observations that were made led to the use not only of quantification of SNP distances, but also the recognition of the genomic locations of SNPs that were shared among cases. Utilization of "shared" SNPs was found to provide power to the use of whole genomics for studying *C. auris* during an outbreak. The concept of shared SNPs allowed rational descriptions and delineations of phylogenetic descendancy. The result was the creation of means to more effectively serve disease controllers and epidemiologists in furtherance of their investigations. While this may not make up for a slow mutation rate associated with the pathogen and thus a lower discriminatory capacity, it does create a means to direct investigators to the most related cases in a rational way. The use of shared SNPs

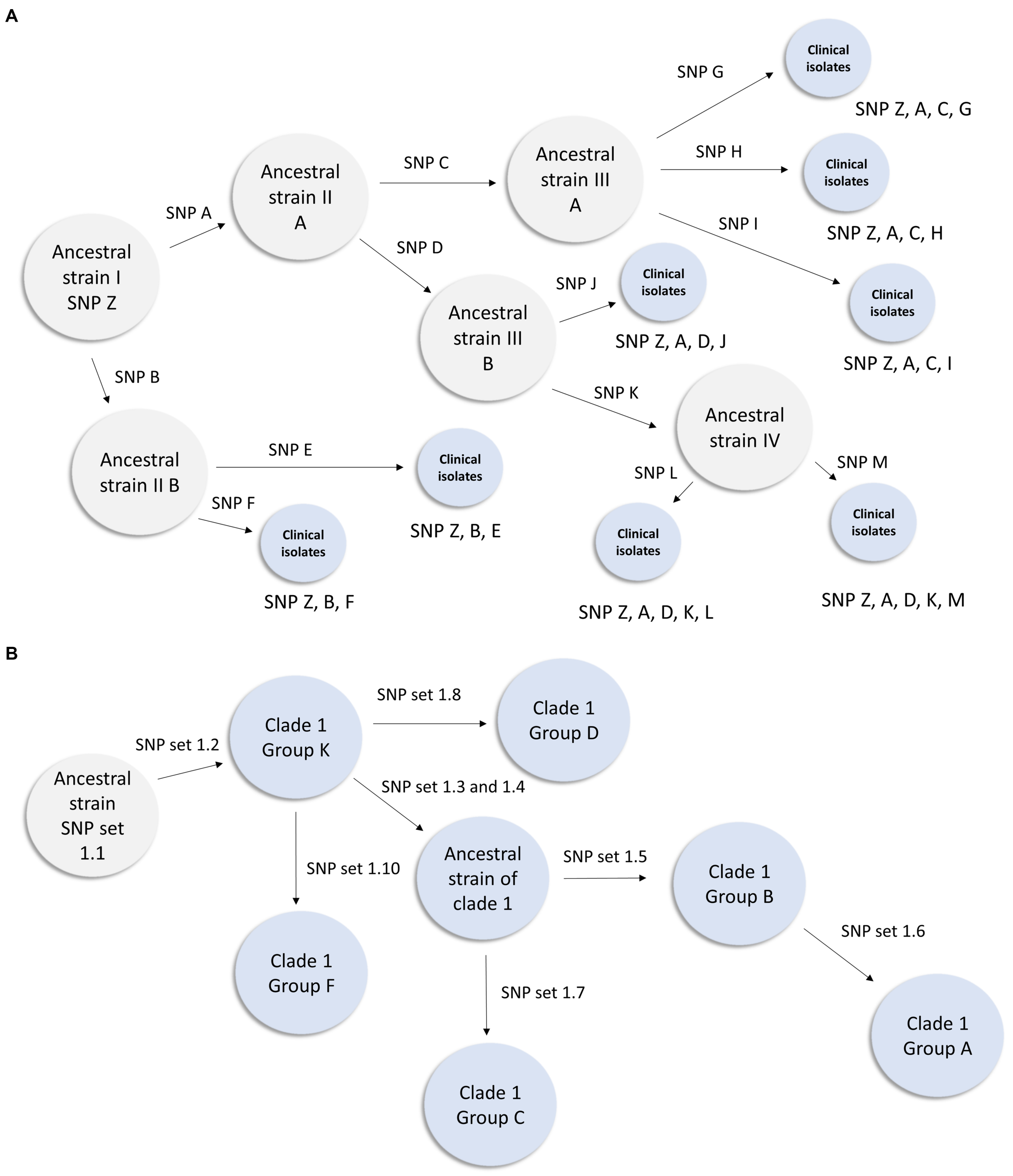


FIGURE 3
Inferred transmission networks based on shared SNP analysis. **(A)** Figure illustrates how theoretical Ancestral strain I (with SNP Z) could evolve during an outbreak. When an isolate acquires a novel SNP or SNPs during this hypothetical outbreak, an arrow is displayed and a designation (such as SNP A) is shown above the arrow. Gray circles represent inferred genotypes based on clinical isolates. Blue circles represent clinical isolates and their known genotypic profile. This figure demonstrates how lineage can be inferred from a genetically diverse set of isolates within a known outbreak. **(B)** Inferred transmission network from the clade 1 southern Nevada outbreak. In this instance we have observed isolates from all intermediates except the most ancestral strain. This representation is derived from the data presented in Table 4.

has been applied to pathogen genomics in many instances but to our knowledge, this is the first use of the concept to assess case relatedness within a *C. auris* outbreak (27–29). A previous study applied whole genome sequencing to a small outbreak of *C. auris* within a hospital (30). Therein it was shown that in fact there was considerable genomic diversity between multiple isolates taken from the same patient, and also taken from different patients who were roomed together and were likely transmission pairs (30).

TABLE 4 Shared SNP analysis of clade 1 isolates.

	SNP Set	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10	1.11	1.12	1.13	1.14	1.15	1.16	1.17	1.18	1.19	Clade 1 Group
	# SNP in Set	6	1	1	1	1	1	4	3	8	1	1	1	2	3	1	1	1	3	2	
SRR23958479	|	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	**A**
SRR23958550	|	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	**A**
SRR19664607	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR20081625	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR23958553	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR23958512	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR23109092	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR19738655	|	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**B**
SRR23958484	|	+	+	+	+	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	**C**
SRR23958472	|	+	+	+	+	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	**C**
SRR23958509	|	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**n/a**
SRR20081622	|	+	+	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	**D**
SRR20081630	|	+	+	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	**D**
SRR23958473	|	+	+	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	**D**
SRR23958536	|	+	+	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	**E**
SRR19738708	|	+	+	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	**E**
SRR20081617	|	+	+	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	**F**
SRR23958480	|	+	+	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	**F**
SRR23958501	|	+	+	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	**G**
SRR23958477	|	+	+	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	**G**
SRR23958478	|	+	+	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	**H**
SRR19738637	|	+	+	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	**H**
SRR19738652	|	+	+	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	**I**
SRR23249014	|	+	+	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	**I**
SRR23249013	|	+	+	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	**J**
SRR19738642	|	+	+	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	**J**
SRR23958528	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958523	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958519	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958514	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958503	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958502	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958500	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958487	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19664611	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19738706	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19738640	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19738712	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR20081637	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958556	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR23958546	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19738698	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**
SRR19738690	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**K**

(Continued)

TABLE 4 (Continued)

	SNP Set	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10	1.11	1.12	1.13	1.14	1.15	1.16	1.17	1.18	1.19	Clade 1 Group
	# SNP in Set	6	1	1	1	1	1	4	3	8	1	1	1	2	3	1	1	1	3	2	
SRR19738686	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	K
SRR19738668	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	K
SRR23109093	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	K
SRR23109093	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	K
SRR23249015	|	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	K
SRR19738677	|	+	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	L
SRR19738650	|	+	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	L
SRR23958539	|	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	n/a
SRR19738715	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	+	M
SRR23109090	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	+	M
SRR23109088	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	–	N
SRR19738688	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	–	N
SRR19664610	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	+	–	O
SRR23958488	|	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	+	–	O

SNP Set refers to a SNP or set of SNPs that are found only in a subset of strains in this analysis. # SNP in Set refers to the number of SNPs in each SNP set (this ranges from 1 SNP to 8 SNPs). The Clade 1 Group refers to assigned designations of all isolates that have the same shared SNP pattern.

TABLE 5 Comparison of SNPs in reference to the CDC Clade 1 reference for two strains from the Nevada outbreak.

Isolate	Number of novel mutations compared to the NV clade 1 core isolates
SRR23137821	2,519
SRR23920687	87

Comparison of SNPs in reference to the CDC Clade 1 reference for two strains from the Nevada outbreak.

Sequencing and genomic analysis provided real benefits to disease controllers and epidemiologists who were investigating these outbreaks in Nevada. It became readily possible to distinguish between ongoing transmission within facilities versus novel introductions into facilities on the basis of shared SNP descendancy. This triggered different strategies and tactics on a facility-by-facility basis which were applied based upon phylogenetic information rather than from best guesses. Lastly, we demonstrated the shared SNP analysis detected two novel clade 1 introductions from outside of southern Nevada. This early detection allowed our public health responders to attempt to contain these new introductions and prevent their establishment in southern Nevada. This was critical because the greater the number of overlapping and ongoing outbreaks a region is experiencing, the more complicated the role of disease control investigators and epidemiologists becomes.

Because multiple tools exist to assess whole genome sequencing of *C. auris*, the work herein rigorously compared and contrasted three such pipelines. Each performed reliably, though specific differences in genome sizes and time-to-answer were found among the three. Notably, pairwise comparisons of SNP distances between fixed sets of isolates across different pipelines revealed that different pipelines will provide different results. This finding indicates that choice and validation of pipelines is not just a matter of formality. As microbiology and bioinformatics continue to merge, it is critical that when new pipelines are constructed that they are validated against existing tools. An ever-increasing number of pipelines does not serve the field of medicine or public health if the pipelines are not clearly assessed from a quality assurance perspective. Much work lies ahead for standards, consultation and accreditation agencies associated with diagnostic science, as emerging bioinformatic tools require rigorous assessment. Even when they are not used as diagnostic tools, their use as aids to epidemiology and disease control will trigger enormous shifts in work-time and resources, which are often limited in the public health realm.

Comprehensive and rapid sequencing of cases as described herein has just begun to impact the public health intervention aspect of the outbreak. Affected sites with continuous transmission have sought additional interventions, including novel means of chemical disinfection and the use of (PCR) screening tests for incoming patients and employees. Unlike the use of sequencing for other hospital acquired infections (e.g., CRE) the comprehensive use of sequencing as shown herein has laid the groundwork for training and familiarization with the use of genomic sequence intelligence. The intense sequencing has additionally led to a highly sophisticated and detailed description of the outbreaks which has gained the attention of elected public servants in the state who have sought additional resources for approaching the outbreak. Additionally, sequencing and analysis have also provided gravitas to the successful actualization of *C. auris* as a reportable entity in Nevada.

This study possesses unique strengths. It included a large number of isolates, collected prospectively in the course of major outbreaks.

The study included analysis of two simultaneous, genomically distinct outbreaks (clade I and clade III), which on the surface resembled a singular outbreak. Additionally, the study compares different tools/pipelines rather than merely showing the construction and functionality of one alone.

This study possesses weaknesses of note. While a high number of isolates associated with a large outbreak were assessed, there are significant gaps in the information that matches epidemiologic data to sequencing data. It is difficult to say with certainty that genomic relatedness ascertained herein is guaranteed to be meaningful from the disease control perspective, without more data. Additionally, not all laboratories or public health jurisdictions could necessarily repeat what was performed herein, as massive resources were necessarily harnessed to generate the granularity of intelligence described.

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

AG generated data, performed analysis, wrote sections of the manuscript, and helped in revisions. FA and MS created bioinformatic pipelines, performed analysis, wrote sections of the manuscript, and helped in revisions. LM generated a figure and helped in revisions. DS, CH, PD, ES, and CN generated data for the paper and helped in revisions. KL and JS supported and funded the creation of bioinformatic pipelines and helped in revisions. SVH provided administrative supervision of the work. MP conceived of and authored portions of the manuscript and provided review. DH conceived of the projects, generated data, performed analysis, wrote the manuscript, and helped in revisions. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Centers for Disease Control and Prevention (CDC) Surveillance for Healthcare-Associated and Resistant Pathogens (SHARP); and CDC Epidemiology and Laboratory Capacity (ELC) funding.

Acknowledgments

NB also thanks Celina Morrison and Laurel Stringer for help with data collection. We would like to acknowledge Kimisha Causey for disease control investigation support in the course of this work.

Conflict of interest

Authors FA, MS, KL and JS were employed by Theiagen Consulting LLC.

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1198189/full#supplementary-material

References

1. Satoh K, Makimura K, Hasumi Y, Nishiyama Y, Uchida K, Yamaguchi H. Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiol Immunol.* (2009) 53:41–4. doi: 10.1111/j.1348-0421.2008.00083.x

2. Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, et al. Simultaneous emergence of multidrug-re-sistant Candida auris on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. *Clin Infect Dis.* (2017) 64:134–40. doi: 10.1093/cid/ciw691

3. Bagal UR, Phan J, Welsh RM, Misas E, Wagner D, Gade L, et al. MycoSNP: a portable workflow for performing whole-genome sequencing analysis of Candida auris. *Methods Mol Biol.* (2022) 2517:215–28. doi: 10.1007/978-1-0716-2417-3_17

4. CDC. Tracking Candida auris (2022) Available at: https://www.cdc.gov/fungal/candida-auris/tracking-c-auris.html

5. Chow NA, De Groot T, Badali H, Abastabar M, Chiller TM, Meis JF. Potential fifth clade of Candida auris, Iran, 2018. *Emerg Infect Dis.* (2019) 25:1780–1. doi: 10.3201/eid2509.190686

6. Kenters N, Kiernan M, Chowdhary A, Denning DW, Pemán J, Saris K, et al. Control of Candida auris in healthcare institutions: outcome of an International Society for Antimicrobial Chemotherapy expert meeting. *Int J Antimicrob Agents.* (2019) 54:400–6. doi: 10.1016/j.ijantimicag.2019.08.013

7. Ong CW, Chen SC, Clark JE, Halliday CL, Kidd SE, Marriott DJ, et al. Australian and New Zealand mycoses interest group (ANZMIG), healthcare infection control special interest group (HICSIG); both of the Australasian Society for Infectious Diseases (ASID). Diagnosis, management and prevention of Candida auris in hospitals: position statement of the Australasian Society for Infectious Diseases. *Intern Med J.* (2019) 49:1229–43. doi: 10.1111/imj.14612

8. PHE. Candida auris identified in England. *Public Health England.* (2016) Available at: https://www.gov.uk/government/publications/candida-auris-emergence-in-england/candida-auris-identified-in-england

9. ECDC. Candida auris in healthcare settings – Europe. European Centre for Disease. *Prev Control.* (2016) Available at: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/Candida-in-healthcare-settings_19-Dec-2016.pdf

10. CDC. Candida auris clinical update. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), division of foodborne, waterborne, and environmental diseases (DFWED) (2016). Available at: https://www.cdc.gov/fungal/candida-auris/c-auris-alert-09-17.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Ffungal%2Fdiseases%2Fcandidiasis%2Fc-auris-alert-09-17.html

11. Cortegiani A, Misseri G, Fasciana T, Giammanco A, Giarratano A, Chowdhary A. Epidemiology, clinical characteristics, resistance, and treatment of infections by Candida auris. *J Intensive Care.* (2018) 6:1–13. doi: 10.1186/s40560-018-0342-4

12. Biswas C, Wang Q, van Hal SJ, Eyre DW, Hudson B, Halliday CL, et al. Genetic heterogeneity of Australian Candida auris isolates: insights from a nonoutbreak setting using whole-genome sequencing. *Open Forum Infect Dis.* (2020) 7:ofaa158,. doi: 10.1093/ofid/ofaa158

13. Available at: https://github.com/tseemann/nullarbor

14. Ambrosio F, Scribner MR, Wright SM, Otieno J, Gorzalski A, Siao DD, et al. TheiaEuk: a species-agnostic bioinformatics workflow for fungal genomic characterization. *Front Public Health*. (2023)

15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

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

17. Torsten S. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. (2014) 30:2068–9. doi: 10.1093/bioinformatics/btu153

18. Available at: https://github.com/tseemann/snippy

19. Ewels P, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*. (2020) 38:276–8. doi: 10.1038/s41587-020-0439-x

20. Lumpe J, Gumbleton L, Gorzalski A, Libuit K, Varghese V, Lloyd T, et al. GAMBIT (genomic approximation method for bacterial identification and tracking): a methodology to rapidly leverage whole genome sequencing of bacterial isolates for clinical identification. *PLoS One*. (2023) 18:e0277575. doi: 10.1371/journal.pone.0277575

21. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*. (2015) 31:2877–8. doi: 10.1093/bioinformatics/btv271

22. Carroll LM, Wiedmann M, Mukherjee M, Nicholas DC, Mingle LA, Dumas NB, et al. Characterization of emetic and diarrheal *Bacillus cereus* strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and Bioinformatic challenges. *Front Microbiol*. (2019) 10:144. doi: 10.3389/fmicb.2019.00144

23. Edwards HM, Rhodes J. Accounting for the biological complexity of pathogenic Fungi in phylogenetic dating. *J Fungi (Basel)*. (2021) 7:661. doi: 10.3390/jof7080661

24. Ene IV, Farrer R, Hirakawa M, Agwamba K, Cuomo CA, Bennett RJ. Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proc Natl Acad Sci U S A*. (2018) 115:E8688–97. doi: 10.1073/pnas.1806002115

25. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, et al. Genome sequencing of an extended series of NDM-producing *Klebsiella pneumoniae* isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital-associated transmission in an endemic setting. *Antimicrob Agents Chemother*. (2014) 58:7347–57. doi: 10.1128/AAC.03900-14

26. Zhu YO, Siegal M, Hall D, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*. (2014) 111:E2310–8. doi: 10.1073/pnas.1323011111

27. Tan KK, Tan YC, Chang LY, Lee KW, Nore SS, Yee WY, et al. Full genome SNP-based phylogenetic analysis reveals the origin and global spread of *Brucella melitensis*. *BMC Genomics*. (2015) 16:93. doi: 10.1186/s12864-015-1294-x

28. Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternberg SM, Driebe EM, et al. Next-generation sequencing of Coccidioides immitis isolated during cluster investigation. *Emerg Infect Dis*. (2011) 17:227–32. doi: 10.3201/eid1702.100620

29. Reis AC, Salvador LCM, Robbe-Austerman S, Tenreiro R, Botelho A, Albuquerque T, et al. Whole genome sequencing refines knowledge on the population structure of *Mycobacterium bovis* from a multi-host tuberculosis system. *Microorganisms*. (2021) 9:1585. doi: 10.3390/microorganisms9081585

30. Roberts SC, Zembower TR, Ozer EA, Qi C. Genetic evaluation of nosocomial Candida auris transmission. *J Clin Microbiol*. (2021) 59:e02252–20. doi: 10.1128/JCM.02252-20

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 01 August 2023
DOI 10.3389/fpubh.2023.1198213

TheiaEuk: a species-agnostic bioinformatics workflow for fungal genomic characterization

Frank J. Ambrosio III[1†], Michelle R. Scribner[1†], Sage M. Wright[1], James R. Otieno[1], Emma L. Doughty[1], Andrew Gorzalski[2], Danielle Denise Siao[2], Steve Killian[3], Chi Hua[4], Emily Schneider[4], Michael Tran[4], Vici Varghese[3], Kevin G. Libuit[1], Mark Pandori[2,5,6]*, Joel R. Sevinsky[1*‡] and David Hess[2,5*‡]

[1]Theiagen Genomics, Highlands Ranch, CO, United States, [2]Nevada State Public Health Laboratory, Reno, NV, United States, [3]Alameda County Public Health Laboratory, Oakland, CA, United States, [4]Public Health Laboratories, Division of Disease Control and Health Statistics, Washington State Department of Health, Shoreline, WA, United States, [5]Department of Pathology and Laboratory Medicine, Reno School of Medicine, University of Nevada, Reno, NV, United States, [6]Department of Microbiology and Immunology, Reno School of Medicine, University of Nevada, Reno, NV, United States

OPEN ACCESS

EDITED BY
Peera Hemarajata,
Association of Public Health Laboratories,
United States

REVIEWED BY
Emily Hollister,
Diversigen, Inc, United States
Benjamin Von Bredow,
Beaumont Health, United States
Kelsey Florek,
University of Wisconsin-Madison, United States

*CORRESPONDENCE
Mark Pandori
✉ mpandori@med.unr.edu
Joel R. Sevinsky
✉ joel.sevinsky@theiagen.com
David Hess
✉ dhess@med.unr.edu

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

‡These authors have contributed equally to this work and share last authorship

RECEIVED 31 March 2023
ACCEPTED 04 July 2023
PUBLISHED 01 August 2023

CITATION
Ambrosio FJ, Scribner MR, Wright SM, Otieno JR, Doughty EL, Gorzalski A, Siao DD, Killian S, Hua C, Schneider E, Tran M, Varghese V, Libuit KG, Pandori M, Sevinsky JR and Hess D (2023) TheiaEuk: a species-agnostic bioinformatics workflow for fungal genomic characterization.
Front. Public Health 11:1198213.
doi: 10.3389/fpubh.2023.1198213

Introduction: The clinical incidence of antimicrobial-resistant fungal infections has dramatically increased in recent years. Certain fungal pathogens colonize various body cavities, leading to life-threatening bloodstream infections. However, the identification and characterization of fungal isolates in laboratories remain a significant diagnostic challenge in medicine and public health. Whole-genome sequencing provides an unbiased and uniform identification pipeline for fungal pathogens but most bioinformatic analysis pipelines focus on prokaryotic species. To this end, TheiaEuk_Illumina_PE_PHB (TheiaEuk) was designed to focus on genomic analysis specialized to fungal pathogens.

Methods: TheiaEuk was designed using containerized components and written in the workflow description language (WDL) to facilitate deployment on the cloud-based open bioinformatics platform Terra. This species-agnostic workflow enables the analysis of fungal genomes without requiring coding, thereby reducing the entry barrier for laboratory scientists. To demonstrate the usefulness of this pipeline, an ongoing outbreak of *C. auris* in southern Nevada was investigated. We performed whole-genome sequence analysis of 752 new *C. auris* isolates from this outbreak. Furthermore, TheiaEuk was utilized to observe the accumulation of mutations in the FKS1 gene over the course of the outbreak, highlighting the utility of TheiaEuk as a monitor of emerging public health threats when combined with whole-genome sequencing surveillance of fungal pathogens.

Results: A primary result of this work is a curated fungal database containing 5,667 unique genomes representing 245 species. TheiaEuk also incorporates taxon-specific submodules for specific species, including clade-typing for *Candida auris (C. auris)*. In addition, for several fungal species, it performs dynamic reference genome selection and variant calling, reporting mutations found in genes currently associated with antifungal resistance (*FKS1*, *ERG11*, *FUR1*). Using genome assemblies from the ATCC Mycology collection, the taxonomic identification module used by TheiaEuk correctly assigned genomes to the species level in 126/135 (93.3%) instances and to the genus level in 131/135 (97%) of instances, and provided zero false calls. Application of TheiaEuk to actual specimens obtained in the course of work at a local public health laboratory resulted in 13/15 (86.7%) correct calls at the species level, with 2/15 called at the genus level. It made zero incorrect calls. TheiaEuk accurately assessed clade type of *Candida auris* in 297/302 (98.3%) of instances.

Discussion: TheiaEuk demonstrated effectiveness in identifying fungal species from whole genome sequence. It further showed accuracy in both clade-typing of *C. auris* and in the identification of mutations known to associate with drug resistance in that organism.

KEYWORDS

***Candida auris*, epidemiology, whole-genome sequencing, bioinformatics, emerging pathogens**

1. Introduction

Microbial fungal pathogens are a major public health concern estimated to affect over 13 million patients annually, with mortality of over 1 million patients annually (1, 2). Fungal infections are especially problematic for patients with conditions such as HIV/AIDs, chronic obstructive pulmonary disease (COPD), asthma, tuberculosis and patients undergoing cancer treatments. Fungal pathogens remain understudied compared to prokaryotic pathogens and often present difficulties in identification and characterization (3–8).

Antifungal drugs are the primary treatment for pathogenic fungal infections. There are four major classes of antifungal drugs: echinocandins (caspofungin), azoles (fluconazole), polyenes (amphotericin B), and the pyrimidine analogue 5-flucytosine. However, the overuse and misuse of these drugs have led to the emergence of drug-resistant strains of these fungi and increasingly prevalent multi-drug resistant fungal infections (9–11). Given the limited classes of drugs to treat fungal infections, the threat of multidrug resistant fungal infections poses a public health menace (9, 11). These strains are often more difficult to treat, resulting in longer hospital stays, higher healthcare costs, and increased mortality rates. In fact, some studies have shown that mortality rates can be as high as 50% in patients with drug-resistant *Candida albicans* infections (9, 10).

Candida auris is a fungal pathogen that has rapidly emerged as a public health concern. It was originally identified in Japan in 2009, and has since been found in over 30 countries, including the United States (10, 12–14). This organism is particularly concerning because it has demonstrated resistance to multiple antifungal drugs, making treatment of infections challenging. In a study of *C. auris* isolates from multiple continents, fluconazole resistance was detected in 93% of isolates, amphotericin B resistance was detected in 35%, and echinocandin resistance was detected in 7% (13). The scope of antifungal treatment options is limited, making managing infections with *C. auris* difficult (1). The ability to resist treatment combined with the ability to cause invasive infections in patients who are already ill and weakened leads to high *C. auris* mortality (13, 15). This highlights the need for enhanced surveillance methods that detect not only the presence of *C. auris,* but also whether the isolate is part of an ongoing outbreak and what antifungal resistance determinants the isolate may harbor.

While there are numerous other fungal pathogens of public health concern, certain species exist as growing antimicrobial resistance threats. *Aspergillus fumigatus* is a common opportunistic airborne fungal pathogen that can cause serious infections in humans. Resistance to several antifungal drugs, including azoles, has been observed in this fungus (16, 17). *Cryptococcus neoformans* is a fungal pathogen that causes serious infections in individuals with weakened immune systems, and often presents difficulties in infection management due to resistance to several antifungal drugs (18, 19). *C. albicans* is a type of fungus commonly found on the skin and mucous membranes of humans. Although often harmless, it can cause infections in vulnerable individuals, such as those with weakened immune systems, surgical wounds, or indwelling medical devices. In recent years, *C. albicans* has also become a growing public health threat due to its increasing resistance to antifungal drugs (20, 21).

Genomic sequencing is a useful tool for analyzing fungal pathogens for public health investigations (22). By analyzing individual pathogen genomes, researchers can identify the species responsible for a patient infection, sub-type the organism, and detect mutations that are associated with resistance to antifungal medicines. For this to be realized, accessible and easy-to-use bioinformatic pipelines for genomic fungal analysis must be developed and deployed to the public health community. To this end, we developed TheiaEuk, a pipeline that performs genome assembly and taxonomic identification of 245 fungal species across 138 genera from FASTQ files generated by whole-genome sequencing. Following taxonomic identification, species-specific analyses are automatically launched. For example, when *C. auris* is detected, clade designation and mutations that are likely to result in antifungal resistance are automatically reported. Lastly, genome assemblies produced by the TheiaEuk pipeline are compatible with several tools for downstream phylogenetic analysis especially when accessed in the Terra platform (23). We demonstrate that the TheiaEuk pipeline provides the bioinformatic tools needed by public health and medical professionals to utilize whole-genome sequencing to characterize and to phylogenetically assess fungal pathogens.

2. Materials and methods

2.1. TheiaEuk pipeline

2.1.1. TheiaEuk implementation

The TheiaEuk workflow was designed to perform *de novo* genome assembly, quality assessment, and genomic characterization of fungal pathogen genomes from paired-end short read sequencing data (see text footnote 1). The workflow is written in the workflow description language (WDL) and as such may be implemented on the browser-based Terra platform (23, 24). The workflow can also be executed from the command line interface using WDL workflow engines such as Cromwell or miniWDL (25, 26). TheiaEuk will process and analyze Illumina paired-end FASTQ inputs using default parameters established for robust fungal pathogen analysis; these parameters can be modified by users from within the graphical user interface of Terra. The workflow utilizes many existing bioinformatics tools as cited in the sections below and produces outputs with industry standard file formats to facilitate downstream analyses. Comparison of TheiaEuk to other pipelines that have been deployed for fungal genome analysis, MycoSNP (27) and Nullarbor (28), was presented in Gorzalski et al. (29). The structure of the pipeline is described below and illustrated in Figure 1.

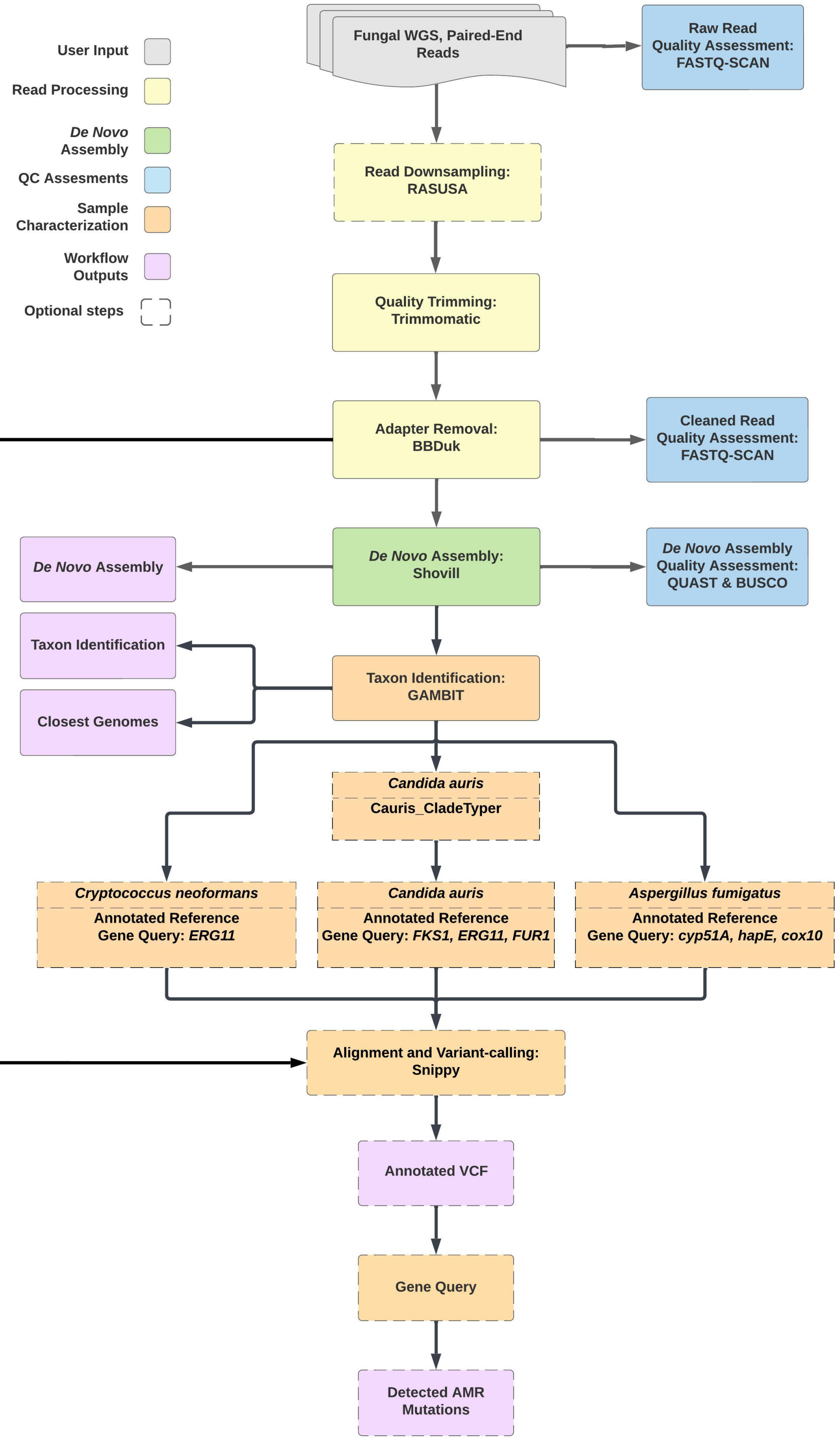

User Input
Read Processing
De Novo Assembly
QC Assesments
Sample Characterization
Workflow Outputs
Optional steps
Fungal WGS, Paired-End Reads
Raw Read Quality Assessment: FASTQ-SCAN
Read Downsampling: RASUSA
Quality Trimming: Trimmomatic
Adapter Removal: BBDuk
Cleaned Read Quality Assessment: FASTQ-SCAN
De Novo Assembly
De Novo Assembly: Shovill
De Novo Assembly Quality Assessment: QUAST & BUSCO
Taxon Identification
Closest Genomes
Taxon Identification: GAMBIT
Candida auris
Cauris_CladeTyper
Cryptococcus neoformans
Annotated Reference Gene Query: ERG11
Candida auris
Annotated Reference Gene Query: FKS1, ERG11, FUR1
Aspergillus fumigatus
Annotated Reference Gene Query: cyp51A, hapE, cox10
Alignment and Variant-calling: Snippy
Annotated VCF
Gene Query
Detected AMR Mutations

FIGURE 1
The TheiaEuk workflow is a species-agnostic bioinformatics pipeline for fungal genome characterization. Input FASTQ files from WGS of fungal pathogens are assessed for quality and *de novo* assembled regardless of species. Taxonomic identification is performed by GAMBIT using a custom

(Continued)

FIGURE 1 (Continued)
fungal database. Taxa-specific sub-workflows for *Candida auris*, *Cryptococcus neoformans*, and *Aspergillus fumigatus* proceed automatically based on the GAMBIT result.

2.1.2. Read trimming and quality control

To avoid errant characterization from poor sequencing data, TheiaEuk performs raw read screening on input FASTQ files to determine whether the workflow will proceed to subsequent analysis or be halted in the event of scarce or problematic input data. This step assesses the number of base pairs, number of reads, and proportion of reads in each input FASTQ file. It also employs MASH sketches to estimate genome size and sequencing depth (30). Samples that pass the initial read screen proceed to an optional step in which reads are randomly subsampled to 150× read depth using RASUSA to conserve computational resources (31). Next, TheiaEuk performs read trimming using Trimmomatic and adapter trimming using BBDuk (32, 33). Read trimming is followed by an additional read screening step to determine if the sequencing data still passes the screening parameters. Samples which meet the parameters proceed automatically to genome assembly.

2.1.3. Genome assembly

TheiaEuk performs *de novo* genome assembly using the Shovill package (34). Shovill is a software package containing several assembly algorithms commonly used for bacterial genome assembly including SPAdes (35) and SKESA (36). SKESA has been set as the default assembler, but the ability to select an alternative assembly program is made available to the user. All the assembly programs within Shovill are designed to assemble haploid genomes, which limits the scope of the pipeline to fungal pathogens with single copies of unpaired chromosomes. Certain downstream modules, particularly GAMBIT due to its *k*-mer-based approach, may be robust to bioinformatics challenges associated with *de novo* assembly of diploid organisms. Nonetheless, these assemblies may be highly fragmented or error prone. Results for diploid organisms must at minimum be assessed with caution and in conjunction with the level of heterozygosity. Following *de novo* assembly, TheiaEuk performs quality assessment of the assembly using QUAST and BUSCO (37, 38).

2.1.4. Taxonomic identification

Following genome assembly, the assembly FASTA files are passed to the Genomic Approximation Method for Bacterial Identification and Tracking (GAMBIT) tool for taxonomic identification (39). GAMBIT infers taxonomy by querying a sample genome against a database of genomes with known taxonomic information and identifying the most similar genome to the query. If the distance between the query genome and the closest genome is within a built-in species threshold, GAMBIT reports the species of the closest genome as the predicted species for the query genome. If not, GAMBIT determines if the query is close enough to be considered a member of the closest genome's genus, otherwise it will not make a taxonomic prediction for the query genome.

Within TheiaEuk, GAMBIT is implemented with the default parameters ($k = 11$ and prefix = ATGAC), and the taxon predicted by GAMBIT is reported as well as the ten closest genomes within the GAMBIT database to the query sample. The only previously published GAMBIT database is exclusive to prokaryotic species, therefore we developed a novel fungal database for identification of fungal pathogens, as described below. This novel fungal database is used by default within TheiaEuk.

2.1.5. Taxa-specific modules (clade typing)

Based on the taxonomic identification made by GAMBIT, TheiaEuk proceeds with taxa-specific modules. For samples identified as *C. auris*, TheiaEuk will perform clade typing using GAMBIT with a custom database consisting of five reference genomes representing the five major clades of *C. auris* (Supplementary Table S1). GAMBIT reports the reference genome that is most similar to the query genome and the associated clade is reported for the sample.

2.1.6. Taxa-specific modules (AMR determinant detection)

For samples identified as *C. auris*, *A. fumigatus*, and *C. neoformans*, TheiaEuk invokes a module which aligns input FASTQ files to a species-appropriate annotated reference genome using Snippy (40). To detect potential antimicrobial resistance determinants, the resulting VCF files may be queried for gene and product names that are associated with antimicrobial resistance following TheiaEuk analysis. Snippy has been used previously to detect mutations in the *FKS1* gene of *Candida* species (41). For *C. auris*, the antimicrobial resistance detection module aligns reads to a clade-specific reference genome and automatically queries the resulting VCF files for three genes associated with antimicrobial resistance (*FKS1*, *ERG11*, *FUR1*). A list of all mutations that have been detected in these select genes are reported to the user. The reference genomes for each *C. auris* clade are indicated in Supplementary Table S1.

2.2. Fungal GAMBIT database creation

In order to infer taxonomic assignments from fungal genomic data, we created a novel fungal GAMBIT database using a similar process as the prokaryotic GAMBIT database (39). The process of creating a GAMBIT database requires the calculation of compressed representations of each genome that will be included in the database, or GAMBIT signatures, which enable the calculation of GAMBIT distances between genomes. In order for GAMBIT to generate a species assignment for a query genome, the distance between the query genome and the closest genome within the database must be below the maximum distance between genomes within that species (species diameter). As such, the GAMBIT database must be curated to ensure that species diameters are non-overlapping and unbiased by mislabeled or poor-quality genomes.

The novel fungal database was created by downloading all the fungal genomes available on GenBank as of 2022-11-30 and curating this list of genomes to exclude poorly represented species and mislabeled genomes. GAMBIT signatures were computed using the same criteria as the most recent GAMBIT bacterial database ($k = 11$ and prefix = ATGAC). For inclusion in the database, species were required to have at least two genomes in GenBank and at least one genome representing the species in RefSeq (42). Subsequently, we curated the database on the basis of the species diameter. Specifically, we computed the GAMBIT diameter of each species and excluded species with either (i) a diameter of zero or (ii) a combination of three or fewer genomes and a diameter greater than 0.75. The database was also manually curated to remove genomes which were clearly highly distant from all other genomes within the species, as these were likely mislabeled on submission.

To establish a set of genomes with non-overlapping species diameters, it was necessary to divide nine species into subspecies groups. In the event that the closest genome in the database to a query genome is a member of a subspecies, GAMBIT will report the parent species as the taxonomic assignment. In addition, two pairs of species were too closely related to distinguish (*Aspergillus flavus/Aspergillus oryzae* and *Aspergillus niger/Aspergillus welwitschiae*), therefore were combined. If the distance between a query genome and the closest genome in the GAMBIT database is greater than the species diameter, GAMBIT checks if the sample is within the genus diameter and attempts to report a genus for the genome. Genus diameters were computed similarly to species diameters, but were additionally curated by lowering the diameter to 95% of the minimum distance between the genus and other genera in the database and to 20% greater than the maximum species diameter of any species within the genus.

Ultimately, 245 fungal species from 138 genera are represented in the fungal database from a total of 5,667 fungal genomes. A table indicating the number of genomes and species diameter for each species represented in the database is indicated in Supplementary Table S2.

2.3. Fungal GAMBIT database validation

2.3.1. GAMBIT versus ANI analysis

Analysis of GAMBIT distances versus average nucleotide identity (ANI) was performed using the GAMBIT distance values computed during the creation of the fungal database for all of the genomes in set 1 and set 2 ($k = 11$ and prefix = ATGAC). Set 1 included all *Candida* genomes within the fungal GAMBIT fungal database and set 2 included a diverse set of genomes across multiple genera. ANI was computed using FastANI (version 1.33) with default parameter values (*k*-mer size 16 and fragment length 3,000) (43). Pairwise comparisons were included in both the statistical analysis and visualizations if the percent of mapped fragments was at least 50%. Figures were generated using scripts adapted from Lumpe et al. using Matplotlib (44, 45).

2.3.2. ATCC mycology genomes

Validation of the fungal GAMBIT database using the ATCC Mycology Collection genomes was performed using the Gambit_Query workflow developed by Theiagen Genomics on Terra.[1] All available fungal genomes were downloaded from the ATCC genome portal on 2023-03-08 (46–48). ATCC genomes downloaded from the ATCC genome portal were used exclusively for testing and were not included in the GAMBIT fungal database. GAMBIT was run with default parameters and we examined the predicted taxon and predicted taxon rank for agreement with the taxonomic annotation from ATCC.

2.3.3. Sequenced isolates from Alameda County

In order to generate a diverse set of fungal genomes for assessing the accuracy of GAMBIT using the fungal database, 19 fungal samples from 18 distinct species were obtained from the Alameda County Public Health Laboratory. Whole genome sequencing of these fungal specimens was performed by the Nevada State Public Health Laboratory through an identical protocol as described below for sequencing of *C. auris* isolates from southern Nevada. The TheiaEuk workflow v1.0.0 was used to run GAMBIT with default parameters on Terra and we compared the predicted taxon from GAMBIT to the taxonomic assignment made using molecular techniques. Whole genome sequencing data for each specimen was submitted to NCBI's Sequencing Read Archive (SRA); accessions are available in Supplementary Table S3.

2.4. Clade typing validation

Within the TheiaEuk pipeline, clade typing of *C. auris* is performed when a sample is predicted to be *C. auris* by GAMBIT. We tested the accuracy of the TheiaEuk clade typing module by querying 302 samples from a published *C. auris* dataset in which clades were assigned (49). Genomes in this dataset were originally derived from multiple studies, with clade type reported by Chow et al. (13, 49–54). Sequencing read data was pulled from NCBI's SRA using the Theiagen Genomics SRA_Fetch workflow[2] and analyzed using TheiaEuk v1.0.0 with default parameters.

2.5. Antimicrobial resistance mutation detection validation

To verify that TheiaEuk reports mutations in antimicrobial resistance genes in samples with known resistance determinants, we identified whole genome sequencing data for 219 *C. auris* samples from published datasets (55–57). FASTQ files for these samples were imported into Terra using the SRA_Fetch workflow and analyzed using TheiaEuk v1.0.0 with the default parameters. The outcome of the TheiaEuk AMR mutation detection module was compared to the known *FKS1* and *ERG11* mutations within each sample.

[1] https://github.com/theiagen/public_health_bioinformatics/blob/PHB-v0.1.0-theiaeuk-manuscript/workflows/standalone_modules/wf_gambit_query.wdl

[2] https://github.com/theiagen/public_health_bioinformatics/blob/PHB-v0.1.0-theiaeuk-manuscript/workflows/utilities/data_import/wf_sra_fetch.wdl

2.6. Southern Nevada *Candida auris* outbreak

2.6.1. Specimen collection

C. auris specimens from an ongoing outbreak in southern Nevada were isolated from clinical samples collected from April 2022 to February 2023. Genomic data from 752 specimens is reported for the first time in this study, but several analyses utilize all sequenced isolates from the southern Nevada outbreak including an additional 209 specimens reported in Gorzalski et al. (29).

2.6.2. Whole genome sequencing

Genomic DNA for sequencing was extracted using a combination of bead-beating (FastPrep-24, MP Biomedicals, Irvine, CA) and magnetic-bead purification (Maxwell RSC 48, Promega, Madison, WI). First, isolates were picked from Sabouraud Dextrose agar plates and mixed with silica beads (Lysing Matrix C, MP Biomedical). Cells were mechanically sheared with 2 cycles at 6.0 m/s for 30 s with a 5 min pause between (FastPrep-24, MP Biomedical). Genomic DNA was isolated using the PureFood Pathogen Kit (Promega) on a Maxwell RSC 48 (Promega) using the manufacturer's protocol. Genomic DNA libraries were prepared using DNA Prep Kit (Illumina, San Diego, CA) using the manufacturer's recommended protocol using a STARlet automated liquid handler (Hamilton Company, Reno, NV). Paired-end sequencing (2× 151) was performed using Illumina's MiniSeq and NovaSeq 6000 to a minimum depth of 35× average coverage. Whole genome sequencing data for these specimens was submitted to NCBI's sequencing read archive (SRA) and accessions are available in Supplementary Table S4. Samples were analyzed using the TheiaEuk workflow v1.0.0 with default parameters on Terra. Analysis of clade assignments and FKS1 mutations among these samples and an additional 209 specimens reported in Gorzalski et al. (29) was visualized using R and RStudio with the tidyverse package (58–60). Twelve samples with either assembly lengths greater than 14 Mbp or BUSCO completeness scores less than 90% were excluded from this analysis as noted in Supplementary Table S4.

2.6.3. Antimicrobial susceptibility testing

C. auris antimicrobial susceptibility testing (AST) was performed using microbroth dilution and predefined gradient of antibiotic concentrations (Etest) methods. A patient isolate was grown on SabDex agar plate and incubated at 30°C in ambient air for 24 h and used to make 0.5 McFarland inoculum suspension in demineralized sterile water. The 0.5 McFarland suspension was measured by spectrophotometer to verify the 0.5 McFarland (80%–82% transmittance). Twenty microliters of 0.5 McFarland suspension were added into 11 mL of RPMI broth tube and 100 µL of the RPMI diluted sample was distributed to each well of a 96-well plate pre-loaded with antibiotics, then incubated along with control plates for 24 h at 35 °C. The same 0.5 McFarland inoculum suspension was used to inoculate a RPMI agar plate using a sterile cotton swab. A single Amphotericin B Etest strip was applied to middle of the agar surface using sterile forceps and incubated along with control plates for 24 h at 35°C. The AST of the microbroth dilution panel was read using a parabolic magnifying mirror to determine the MIC (lowest concentration where there is ≤50% growth compared to growth control well). For the Amphotericin B Etest, MIC was interpreted at a value where there is 100% growth inhibition (number above where the ellipse intercepts Etest strip).

3. Results

3.1. TheiaEuk workflow

In response to an ongoing outbreak of *C. auris* in southern Nevada, Theiagen Genomics and the Nevada State Public Health Laboratory collaborated to develop a bioinformatics pipeline for analyzing *C. auris* WGS data: TheiaEuk. TheiaEuk is a species-agnostic workflow for fungal genome characterization that can be implemented through a graphical user interface using Terra. Briefly, this pipeline quality trims and assesses input paired-end short read sequencing data then creates a *de novo* assembly using the SKESA assembler (Figure 1) (36). Using the genome assembly, species taxon identification is performed by the Genomic Approximation Method for Bacterial Identification and Tracking (GAMBIT) tool. GAMBIT implementation in TheiaEuk uses a novel, curated fungal database containing 5,667 genomes and 245 species. For certain identified taxa, taxa-specific workflows are activated, such as a *C. auris* clade-typing tool and antifungal resistance detection.

3.2. Fungal GAMBIT database validation

GAMBIT was designed for microbial taxonomic identification by querying genome assemblies against a database and assigning taxonomy based on curated diagnostic thresholds (39). The initial GAMBIT database contained only prokaryotic genomes, but nothing precluded the extension of GAMBIT to eukaryotic microbes. Here we describe the development and validation of a novel fungal microbial database using the core GAMBIT logic.

First we demonstrate that eukaryotic microbial isolates have the same relationship as prokaryotes when comparing average nucleotide identity (ANI) versus GAMBIT distance (Figure 2) (39, 43). To this end, two sets of genomes were selected within the fungal database and ANI and GAMBIT distance computations were performed between every pair of genomes within each dataset. These fungal genomes demonstrate the same logarithmic relationship between ANI and GAMBIT distance as prokaryotic genomes (Figure 2A) which suggests that there is no difference between prokaryotic and eukaryotic microbes in terms of identification via GAMBIT. In the first dataset, we examined 318 genomes from the *Candida* genus (Figure 2B). For comparisons where FastANI reported an ANI value and the percent of mapped fragments was greater than 50% (13,389 genome pairs, 26.4% of comparisons), GAMBIT distance and ANI exhibited a Spearman correlation of 97.3%. This analysis was extended to a broader range of eukaryotic microbial species and demonstrated the same relationship with a Spearman correlation of 98.9% for pairwise comparisons where ANI values were reported (970 genome pairs, 12.3% of comparisons) (Figure 2C).

3.3. Validation of the fungal GAMBIT database using ATCC genomes

To assess the accuracy of the fungal GAMBIT database, the taxonomic assignments were validated using two sets of genomes with

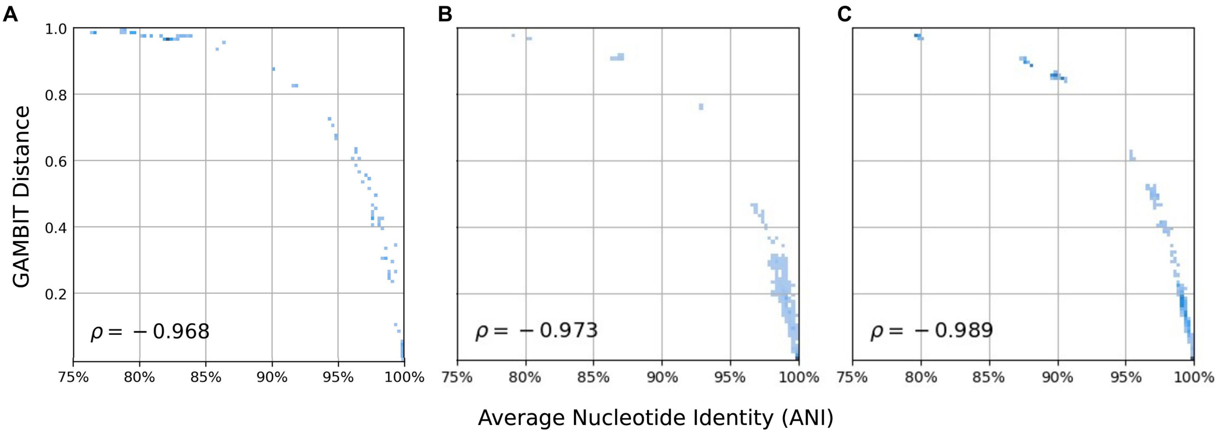


FIGURE 2
Relationship between GAMBIT distance and average nucleotide identity (ANI) for every pair of genomes within three datasets. **(A)** Image is reproduced from (39) and consists of 70 prokaryotic genomes from diverse taxa. This set is shown here to visualize the trend observed in prokaryotic genomes. **(B)** Set 1 consists of 381 *Candida* genomes (Spearman correlation = 97.3%) and **(C)** and set 2 consists of 126 fungal genomes from diverse species (Spearman correlation = 98.9%). Only pairwise comparisons where FastANI reported that greater than 50% of fragments mapped were included in statistical analyses and plotted.

TABLE 1 Fungal GAMBIT database validation using ATCC Mycology genome collection.

		Expected assignment (ATCC)			Total
		Species	Genus	No assignment	
Observed assignment (GAMBIT)	Species	126	0	0	126
	Genus	5	3	0	8
	No assignment	4	30	22	56
	Total	135	33	22	190 Total Genomes

One hundred ninety fungal genomes were analyzed using GAMBIT with the fungal GAMBIT database. Based on the content of the fungal GAMBIT database, query genomes were expected to be identified at the species level, genus level, or not assigned. Correct taxonomic assignments made by GAMBIT for each possible taxonomic rank are shown. GAMBIT made no incorrect taxonomic assignments.

known taxonomic assignments. The first validation was performed using fungal genomes from the ATCC Mycology collection (46, 47). This dataset was selected due to the high level of confidence in the taxonomic assignment of these genomes and includes 190 fungal genomes from 61 genera and 109 species. In total, 135 of the genomes are represented at the species level within the GAMBIT database, 33 are represented only at the genus level, and 22 are not represented at the genus or species level. Of the genomes for which a species prediction was possible (135 genomes), GAMBIT reported the correct species for 126 genomes (Table 1). For the remaining 9 genomes, GAMBIT predicted either the correct genus (5 genomes) or made no taxonomic prediction (4 genomes). For the genomes that were represented only at the genus level (33 genomes), GAMBIT reported the correct genus for 3 genomes and reported no taxonomic assignment for 30 genomes. Finally, for the genomes that were not represented at the genus or species level within the GAMBIT database (22 genomes), GAMBIT made no taxonomic predictions, as expected. GAMBIT taxonomic assignment for each genome is indicated in Supplementary Table S5.

The data demonstrated that using the developed fungal database, GAMBIT reported either an accurate taxonomic assignment or no taxonomic assignment for all 190 genomes examined. Given the underrepresentation of high-quality fungal genomes in public repositories, the GAMBIT database is designed to perform taxonomic identification conservatively. Consequently, the majority of taxonomic assignments were at the lowest possible taxonomic rank (126/135 possible species assignments, 3/33 genus assignments), but 39 genomes were assigned to either a higher taxonomic rank or received no taxonomic assignment.

3.4. Validation of the fungal GAMBIT database using sequenced samples

Given the relative scarcity of fungal genomes available for validating the fungal GAMBIT database, the Nevada State Public Health Laboratory obtained 19 fungal samples from the Alameda County Public Health Laboratory and subjected them to whole genome sequencing. The samples represented 18 distinct fungal species according to previous reference laboratory biochemical and molecular laboratory techniques including *Aspergillus, Candida, Clavispora, Coccidioides, Cryptococcus, Kluyveromyces, Pichia, Trichophyton*, and *Yarrowia* species (Table 2). Sequencing data was analyzed using TheiaEuk with the fungal GAMBIT database to assess the accuracy of GAMBIT taxonomic identification. One sample did not produce quality sequencing data for successful completion of

TheiaEuk (*A. flavus*). Of the remaining 17 species, 14 were represented at the species level within the fungal GAMBIT database, 2 at the genus level only (*Fusarium* of undetermined species and *Candida metapsilosis*), and 1 was not represented (*Trichophyton mentagrophytes*). Of the 15 samples where species assignments were possible, 13 were identified correctly at the species level and 2 were identified correctly at the genus level. Of the 2 samples where genus-level only assignments were possible, 1 was assigned the correct genus and 1 received no assignment. The sample that was not represented in the database received no assignment, as expected. Therefore, both validations of the fungal GAMBIT database demonstrated exclusively accurate taxonomic assignments, often at the lowest taxonomic level possible.

3.5. Clade typing validation

TheiaEuk performs clade typing on genomes that are identified as *C. auris* by GAMBIT using the clade-typer module (Materials and Methods). To validate this functionality, 302 samples with determined clade types from published datasets were compared against the results from TheiaEuk (Table 3) (49). These samples represented four of the five *C. auris* clades (clade I: 126 samples, clade II: 5 samples, clade III: 51 samples, clade IV: 120 samples). All clade assignments made by TheiaEuk were found to match the previously published clade assignments except for one sample which was assigned to clade I despite being previously described as clade III. This genome (strain B16401) was also previously assigned to clade I by another genomic analysis approach, suggesting that the clade identity is controversial for this strain (41). Four samples were not assigned to clades because GAMBIT failed to confidently assign the sample as *C. auris*. Clade typing outcomes for each specimen are available in Supplementary Table S6. TheiaEuk performed accurate clade assignment in 99.6% of cases and therefore enables rapid determination of sample clade without the need for other phylogenetic analysis.

3.6. Antimicrobial resistance determinant detection validation

TheiaEuk detects mutations in select antimicrobial resistance genes by aligning reads to a *C. auris* clade-specific reference genome

TABLE 2 Fungal GAMBIT database validation using genomes obtained from the Alameda County Public Health Laboratory and sequenced by the Nevada State Public Health Laboratory.

NCBI organism name	Expected GAMBIT genus assignment	Expected gambit species assignment	Observed GAMBIT genus assignment	Observed gambit species assignment	Identification method or isolate source
Aspergillus terreus	*Aspergillus*	*terreus*	*Aspergillus*	*terreus*	MALDI-TOF at MDL
Candida albicans	*Candida*	*albicans*	*Candida*	*albicans*	ATCC 14053
Candida auris	*Candida*	*auris*	*Candida*	*auris*	CDC B11903
Candida dubliniensis	*Candida*	*dubliniensis*	*Candida*	NA	Unknown
Candida glabrata	*Candida*	*glabrata*	*Candida*	*glabrata*	ATCC 2001
Candida metapsilosis	*Candida*	NA	NA	NA	MALDI-TOF at MDL
Candida parapsilosis	*Candida*	*parapsilosis*	*Candida*	*parapsilosis*	MALDI-TOF at MDL
Candida tropicalis	*Candida*	*tropicalis*	*Candida*	*tropicalis*	CAP B-36-90
Clavispora lusitaniae	*Clavispora*	*lusitaniae*	*Clavispora*	*lusitaniae*	CAP F-15-00
Coccidioides immitis	*Coccidioides*	*immitis*	*Coccidioides*	*immitis*	Coccidioides real-time PCR at Reference Lab
Coccidioides immitis	*Coccidioides*	*immitis*	*Coccidioides*	*immitis*	Coccidioides real-time PCR at Reference Lab
Cryptococcus gattii VGI	*Cryptococcus*	*gattii*	*Cryptococcus*	*gattii*	ATCC MYA 4560
Cryptococcus neoformans	*Cryptococcus*	*neoformans*	*Cryptococcus*	*neoformans*	ATCC 204092
Fusarium sp.	*Fusarium*	NA	*Fusarium*	NA	Morphology
Kluyveromyces marxianus	*Kluyveromyces*	*marxianus*	*Kluyveromyces*	NA	ATCC 2512
Pichia kudriavzevii	*Pichia*	*kudriavzevii*	*Pichia*	*kudriavzevii*	CAP B-24-92
Trichophyton mentagrophytes	NA	NA	NA	NA	ATCC 9533
Yarrowia lipolytica	*Yarrowia*	*lipolytica*	*Yarrowia*	*lipolytica*	MALDI-TOF at MDL

Expected genus or species assignments were determined by the reference laboratory using the molecular or biochemical approaches indicated. The NCBI organism name column indicates the known taxonomic information about the sample based on molecular or biochemical approaches. The expected GAMBIT genus assignment and expected gambit species assignment columns indicate the expected taxonomic assignment by GAMBIT based on the representation of that taxon within the GAMBIT database. An "NA" is shown if either the genus or species is missing from the database. The observed GAMBIT genus assignment and observed gambit species assignment columns indicate the actual taxonomic assignment by GAMBIT. An "NA" is shown if GAMBIT did not report an assignment at that taxonomic level.

TABLE 3 Clade typing validation using 302 *C. auris* samples spanning four major clades.

	Clade from publication				
Clade-typer results	Total: 302	Clade I	Clade II	Clade III	Clade IV
	Clade I	123	0	1	0
	Clade II	0	5	0	0
	Clade III	0	0	50	0
	Clade IV	0	0	0	119
	Clade-typer skipped	3	0	0	1

Cladetyper results compared to the clades assigned by the original publication (49). The clades assigned by the clade-typer module in TheiaEuk are along the left axis and the clades from the original publication are across the top. Samples that were not successfully assigned to the *C. auris* taxa by GAMBIT were skipped by the clade-typer module. Only one sample produced a discordant result between the clade-typer result (clade I) and the result reported in the original publication (clade III).

TABLE 4 TheiaEuk accurately identified mutations in *FKS1* (top) and *ERG11* (bottom) for 219 *C. auris* genomes from published datasets.

		Expected	
		FKS1 mutation	No *FKS1* mutation
Observed	*FKS1* mutation	44	0
	No *FKS1* mutation	0	175

		Expected	
		ERG11 mutation	No *ERG11* mutation
Observed	*ERG11* Mutation	161	0
	No *ERG11* Mutation	0	58

Samples spanned four *C. auris* clades: clade I (33 samples), clade II (7 samples), clade III (94 samples), and clade IV (85 samples). The number of expected missense, stop codon, and indel mutations detected in *FKS1* and *ERG11* based on the mutations reported in the original publication was compared to the observed number of mutations in these genes reported by TheiaEuk. The default clade III and IV reference genomes in TheiaEuk include known ERG11 mutations, therefore detection of no variant at that site by TheiaEuk was interpreted as agreement with the original publication.

and querying the resulting variant-calling output for associated gene and product names. We sought to verify that TheiaEuk reports mutations in genes associated with antimicrobial resistance from genomic data with known mutation status. To this end, three published datasets with genomic data spanning four *C. auris* clades were identified in which presence or absence of *FKS1* and *ERG11* mutations was noted (55–57). The genomic data was analyzed using TheiaEuk and determined that TheiaEuk correctly identified all known mutations in *FKS1* and *ERG11* for 219 samples (Table 4, results from each sample are available in Supplementary Table S7). Because TheiaEuk reports these mutations from variant-calling data, the choice of reference genome impacts the mutations reported by TheiaEuk. It is observed that the default clade III reference genome in TheiaEuk incorporates a known azole resistance mutation: *ERG11* V125A/F126L (56). Likewise, the clade IV reference genome incorporates the *ERG11* Y132F mutation (61).

3.7. Implementation of TheiaEuk for the southern Nevada outbreak

Since its development, TheiaEuk has been used to analyze 961 *C. auris* isolates from an ongoing outbreak in southern Nevada. Genomic and phylogenetic analysis of the first 209 samples were reported in Gorzalski et al. (29) and the remaining 752 samples are reported for the first time in this study. These 752 specimens were isolated from samples obtained from either patients presenting with symptoms or through screening of long-term care patients between April 2022 to February 2023. Several medical facilities used the Nevada State Public Health Laboratory for routine screening of *C. auris*. Culturing of all PCR positive samples was attempted with sequencing performed on all culture positive specimens. All samples were identified as *C. auris* by TheiaEuk. Twelve samples were excluded from subsequent analysis due to low genome quality; the remainder were assigned to either clade I ($n = 157$) or clade III ($n = 583$). These data represent an ongoing outbreak; the rapid ability to distinguish which isolates belong to the two major outbreaks and which isolates are part of new introductions based on whole-genome sequencing demonstrates the utility of TheiaEuk as a front-line analysis tool for fungal pathogens.

3.8. Detection of antimicrobial resistance determinants in southern Nevada outbreak

The TheiaEuk pipeline enables monitoring of mutations in genes associated with echinocandin resistance, particularly *FKS1*. The relevance of this analysis in the southern Nevada outbreak was examined by two methods. Firstly, the accumulation of *FKS1* mutations over time was examined during the outbreak using data from this work and Gorzalski et al. (Figure 3) (29). These mutations occur in strains that share the complete genetic background of non-*FKS1* mutant isolates in the Nevada outbreak. Thus, the most parsimonious explanation for the occurrence of *FKS1* mutations is that they evolved during the outbreak, suggesting that they are in response to the treatment by the frontline antifungals for *C. auris* which are all in the echinocandin class. Mutations in *FKS1* were detected in 18 out of 949 samples throughout the outbreak and were found to represent 7 distinct amino acid substitutions: Ser639Phe, Leu640Val, Arg641Gly, Arg641Ser, Asp642Tyr, Leu686Phe, and Ile1361Thr.

Secondly, the MIC data for six antifungals that were available for isolates in this dataset were examined. The data was parsed based on presence or absence of *FKS1* mutations (Figure 4). Among the six antifungals, there are three echinocandins: anidulafungin, caspofungin and micafungin. Isolates with *FKS1* mutations exhibit a significantly reduced susceptibility to echinocandins relative to isolates without *FKS1* mutations (Wilcoxon rank sum test with continuity correction: anidulafungin value of $p = 0.0004511$, caspofungin p-value $= 0.000576$, micafungin p-value $= 0.001556$). Reduced susceptibility to azoles was also observed for isolates with *FKS1* mutations to a lesser extent and this trend was significant in two drugs (Wilcoxon rank sum test with continuity correction: isavuconazole p-value $= 0.024270.02203$, itraconazole p-value $= 0.009552$, posaconazole p-value $= 0.05928$). While FKS1 mutations were correlated with reduced susceptibility to azoles, it is unlikely that they were responsible for the reduced

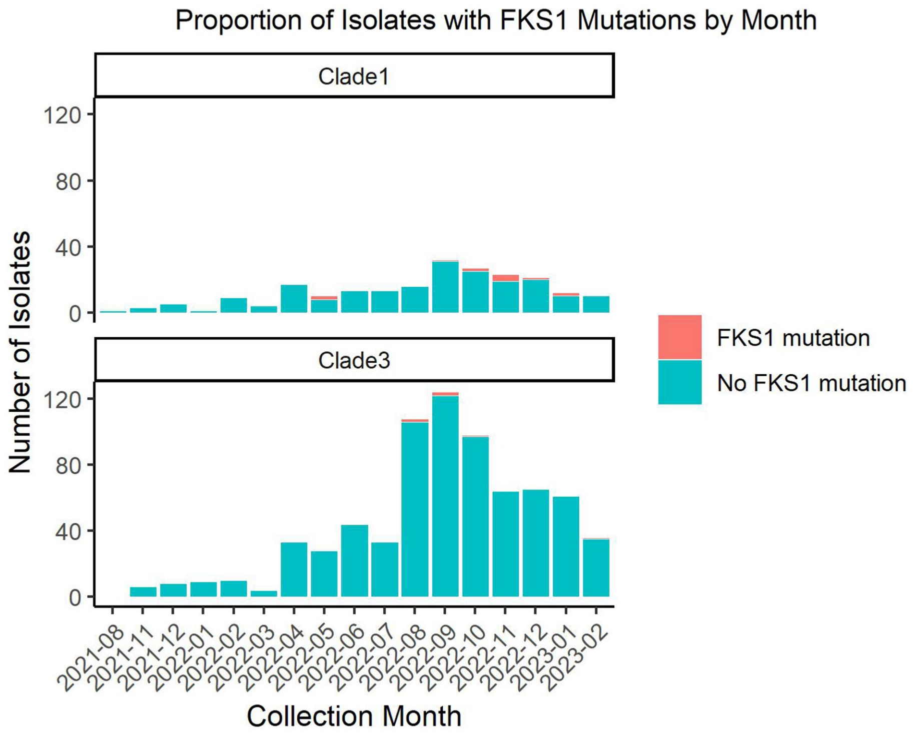


FIGURE 3
Number of southern Nevada *C. auris* isolates with and without *FKS1* mutations by month. Nine hundred forty-nine *C. auris* isolates from southern Nevada were analyzed using TheiaEuk for presence or absence of *FKS1* mutations. This graph splits the isolates between clade I and clade III representing the two major outbreaks in southern Nevada. The data is represented by month with the number of isolates with the wild-type *FKS1* sequence shown in teal and the number of isolates with a mutant *FKS1* sequence shown in orange. This figure excludes one sample collected in January of 2020 which precedes the ongoing outbreaks.

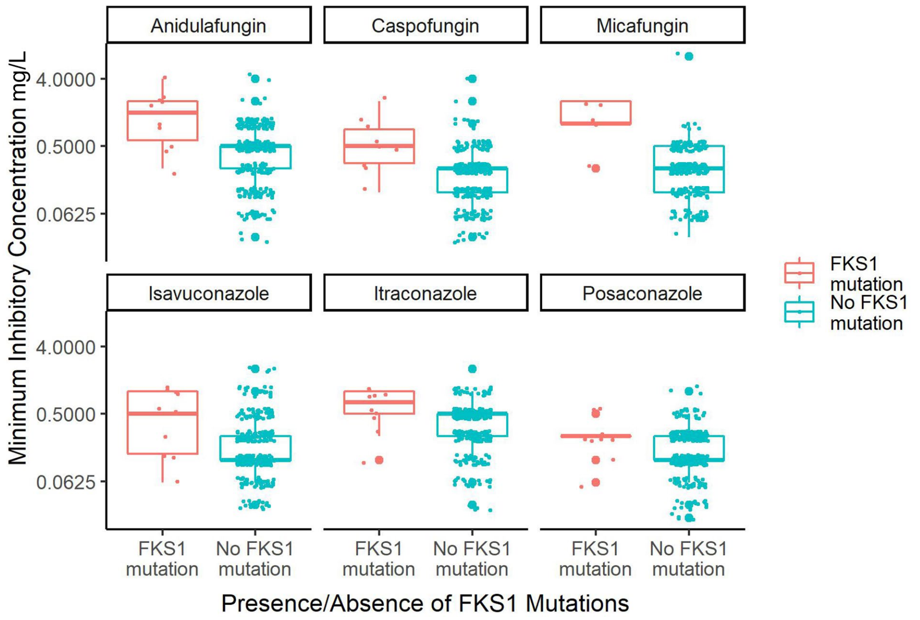


FIGURE 4
Box plot of MIC data for the six major antifungals that treat *C. auris* based on *FKS1* mutation status. Plotted are the 326 isolates from the southern Nevada *C. auris* outbreak that have MIC data for any of six antifungals. Only 247 isolates have MIC data for micafungin, otherwise *n* = 326 for all other drugs. The orange box plots indicate isolates containing *FKS1* mutations and the teal plots indicate isolates containing *FKS1* wild-type sequence. Small dots indicate individual sample MIC measurements whereas large dots indicate outlying data points of the boxplot.

susceptibility given the distinct mechanisms of action of echinocandins and azoles.

4. Discussion

Fungal diseases represent a major threat to public health as evidenced by increasing mortality rates in recent years (62). However, these eukaryotic agents have not incurred the same focus of prokaryotic pathogens, especially in the realm of whole-genome sequence identification and surveillance. This is likely due to the complex nature of their laboratory diagnoses, and the relative paucity of genomic tools to assess them (3–8). The introduction of TheiaEuk provides a platform to utilize whole-genome sequencing of fungal microbial pathogens in both the research and clinical setting.

A novel contribution of this work is the development and assessment of a fungal taxonomic identification process from WGS data. The primary identification engine (GAMBIT) has been utilized in a CLIA regulatory environment to report clinical diagnostic identifications of prokaryotic pathogens (39). Here we extended the same logic to fungal pathogens and laid the groundwork for a similar validation. This is important for clinical laboratories as fungal pathogens often possess complex and ambiguous biochemical profiles that often result in identifications at only the genus level. Moreover, the expertise in mycology to make routine laboratory diagnosis is waning (63). Creating a fungal identification pipeline using whole genome sequencing that will be implemented in a regulatory environment should increase the number of clinically relevant fungal genomes that are produced by public health laboratories and other health care providers (39). This will allow the initial fungal database presented here to be updated and extended to additional fungal species, thus increasing impact.

The regular incorporation of whole genome sequencing to fungal pathogen surveillance provides not only robust taxonomic identification but additional insights regarding genetic relatedness. For example, the use of TheiaEuk in the ongoing *C. auris* outbreak in southern Nevada demonstrated that specimens collected during the same time period represented distinct introductions because it revealed that samples were from two different clades. Also, while the TheiaEuk pipeline does not directly produce phylogenetic trees from specimen sets, the output files generated by the workflow are compatible with numerous downstream tools for more granular phylogenetic analysis. Examples include the kSNP3 workflow and MashTree workflow, both of which are open source and available for analysis using Terra (30, 64, 65). Through these subsequent analyses, transmission networks among fungal pathogens may be discerned.

Examination of the southern Nevada *C. auris* outbreak by TheiaEuk also reveals the necessity of pipelines like the one described for detection of antimicrobial resistance determinants. Currently, there are three classes of antifungals that can treat *C. auris*. Yet, most *C. auris* strains (93%) are resistant to fluconazole, and another 35% are resistant to AmpB (13). This leaves echinocandins as the major frontline defense to *C. auris*. Given that *C. auris* forms biofilms on both biotic and abiotic surfaces, exists asymptomatically on colonized patients, carries drug resistance, and poses potential lethal consequences upon septic infection, *C. auris* presents a real threat to our health care system (66). This threat is amplified if echinocandin resistant isolates become endemic to communities. The ability to detect and to take disease control action on isolates of *C. auris* that have mutations in *FKS1* that correlate with decreased susceptibility to echinocandins is critical to mitigate this new threat. Unfortunately, current phenotype-based systems that assess for decreased susceptibility rely on centralized services where isolates of interest are sent, cultured, then grown and tested against a series of antifungals. This is followed by the reporting of data in a systematic form which often results in a considerable turnaround time to inform health care providers that they have a resistant or decreased susceptibility isolate of *C. auris*. This lag may prevent the most effective actions from being taken to control these potential threat organisms. Whole-genome sequencing and the detection of *FKS1* mutations decrease this timeline significantly and provide a method for disease control investigators to stay ahead of echinocandin resistant strains of *C. auris*.

An often overlooked but increasingly important aspect of bioinformatics tools is the need to be accessible to the broader scientific community, not just bioinformaticians. Innovative tools conceived and developed within the disease pillars of academic and government laboratories are often inaccessible to the average public health scientist with no training, experience, or resources in command line bioinformatics. To this end, we share the same philosophy as Black et al. in their recommendations for supporting open pathogen genomic analysis in public health (67). TheiaEuk was intentionally developed from the beginning to be (1) reproducible in the way it implements containerization, versioning, workflow management, and auditability, (2) scalable in the utilization of cloud resources, and (3) deployable within hours using the open bioinformatics platform Terra for workflow registry and web portal accessibility. This open bioinformatics platform will then bridge across all disease pillars, where specialty tools designed by disease experts will be accessed and utilized in a common, open environment. This is particularly important for public health laboratories whose pathogen genomic outbreak investigations cover the full spectrum of human and animal pathogens. In addition to accessibility the ability to validate workflows for public health use is vital, something not often encountered in research environments but critical for our public health system. Here, again, the use of the open bioinformatics platform Terra, with the ability to version, audit, and validate every workflow, meets the needs of public health scientists, both nationally and internationally.

Data availability statement

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

Author contributions

FA and MS created bioinformatics pipelines, performed analysis, wrote sections of the manuscript, and helped in revisions. SW, JO, and ED created bioinformatics pipelines and helped in revisions. AG generated data for the paper, performed analysis, wrote sections of the manuscript, and helped in revisions. DS, SK, CH, ES, and VV generated data for the paper and helped in revisions. KL supported and funded the creation of bioinformatic pipelines, created bioinformatics pipelines, and helped in revisions. MP conceived of the projects, performed analysis, wrote sections of the manuscript, and helped in revisions. JS supported and funded the creation of bioinformatic pipelines, wrote sections of the manuscript, and helped in revisions. DH conceived of the projects, generated data, performed analysis, wrote sections of the manuscript, and helped in revisions. All authors contributed to the article and approved the submitted version.

Funding

This publication was supported by the Nevada State Department of Health and Human Services through Grant Number NU50CK000560 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department nor the Centers for Disease Control and Prevention.

Acknowledgments

The authors thank Phillip Dykema at Washington State Department of Health for his role in providing antifungal susceptibility data to this project and we also thank Jared Lumpe for demonstrating the process for GAMBIT database creation.

Conflict of interest

FA, MS, SW, JO, ED, KL, and JS were employed by Theiagen Genomics.

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1198213/full#supplementary-material

References

1. Bongomin F, Gago S, Oladele RO, Denning DW. Global and multi-national prevalence of fungal diseases-estimate precision. *J Fungi*. (2017) 3:57. doi: 10.3390/jof3040057

2. Rayens E, Norris KA. Prevalence and healthcare burden of fungal infections in the United States, 2018. Open forum. *Infect Dis*. (2022) 9:ofab593. doi: 10.1093/ofid/ofab593

3. Brown GD, Denning DW, Gow NAR, Levitz SM, Netea MG, White TC. Hidden killers: human fungal infections. *Sci Transl Med*. (2012) 4:165rv13. doi: 10.1126/scitranslmed.3004404

4. Denning DW. Minimizing fungal disease deaths will allow the UNAIDS target of reducing annual AIDS deaths below 500 000 by 2020 to be realized. *Philos Trans R Soc Lond Ser B Biol Sci*. (2016) 371:20150468. doi: 10.1098/rstb.2015.0468

5. Armstrong-James D, Meintjes G, Brown GD. A neglected epidemic: fungal infections in HIV/AIDS. *Trends Microbiol*. (2014) 22:120–7. doi: 10.1016/j.tim.2014.01.001

6. Guinea J, Torres-Narbona M, Gijón P, Muñoz P, Pozo F, Peláez T, et al. Pulmonary aspergillosis in patients with chronic obstructive pulmonary disease: incidence, risk factors, and outcome. *Clin Microbiol Infect*. (2010) 16:870–7. doi: 10.1111/j.1469-0691.2009.03015.x

7. Limper AH, Adenis A, Le T, Harrison TS. Fungal infections in HIV/AIDS. *Lancet Infect Dis*. (2017) 17:e334–43. doi: 10.1016/S1473-3099(17)30303-1

8. Marr KA, Carter RA, Boeckh M, Martin P, Corey L. Invasive aspergillosis in allogeneic stem cell transplant recipients: changes in epidemiology and risk factors. *Blood*. (2002) 100:4358–66. doi: 10.1182/blood-2002-05-1496

9. Fisher MC, Alastruey-Izquierdo A, Berman J, Bicanic T, Bignell EM, Bowyer P, et al. Tackling the emerging threat of antifungal resistance to human health. *Nat Rev Microbiol*. (2022) 20:557–71. doi: 10.1038/s41579-022-00720-1

10. Forsberg K, Woodworth K, Walters M, Berkow EL, Jackson B, Chiller T, et al. *Candida auris*: the recent emergence of a multidrug-resistant fungal pathogen. *Med Mycol*. (2019) 57:1–12. doi: 10.1093/mmy/myy054

11. Hendrickson JA, Hu C, Aitken SL, Beyda N. Antifungal resistance: a concerning trend for the present and future. *Curr Infect Dis Rep*. (2019) 21:47. doi: 10.1007/s11908-019-0702-9

12. Chen J, Tian S, Han X, Chu Y, Wang Q, Zhou B, et al. Is the superbug fungus really so scary? A systematic review and meta-analysis of global epidemiology and mortality of *Candida auris*. *BMC Infect Dis*. (2020) 20:827. doi: 10.1186/s12879-020-05543-0

13. Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, et al. Simultaneous emergence of multidrug-resistant *Candida auris* on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. *Clin Infect Dis*. (2017) 64:134–40. doi: 10.1093/cid/ciw691

14. Satoh K, Makimura K, Hasumi Y, Nishiyama Y, Uchida K, Yamaguchi H. *Candida auris* sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiol Immunol*. (2009) 53:41–4. doi: 10.1111/j.1348-0421.2008.00083.x

15. Chowdhary A, Prakash A, Sharma C, Kordalewska M, Kumar A, Sarma S, et al. A multicentre study of antifungal susceptibility patterns among 350 *Candida auris* isolates (2009–17) in India: role of the ERG11 and FKS1 genes in azole and echinocandin resistance. *J Antimicrob Chemother*. (2018) 73:891–9. doi: 10.1093/jac/dkx480

16. Berger S, El Chazli Y, Babu AF, Coste AT. Azole resistance in *Aspergillus fumigatus*: a consequence of antifungal use in agriculture? *Front Microbiol*. (2017) 8:1024. doi: 10.3389/fmicb.2017.01024

17. Scorzoni L, de Paula E, Silva ACA, Marcos CM, Assato PA, de Melo WCMA, et al. Antifungal therapy: new advances in the understanding and treatment of mycosis. *Front Microbiol*. (2017) 8:36. doi: 10.3389/fmicb.2017.00036

18. Elsegeiny W, Marr KA, Williamson PR. Immunology of cryptococcal infections: developing a rational approach to patient therapy. *Front Immunol*. (2018) 9:651. doi: 10.3389/fimmu.2018.00651

19. Zafar H, Altamirano S, Ballou ER, Nielsen K. A titanic drug resistance threat in *Cryptococcus neoformans*. *Curr Opin Microbiol*. (2019) 52:158–64. doi: 10.1016/j.mib.2019.11.001

20. Thatchanamoorthy N, Rukumani Devi V, Chandramathi S, Tay ST. *Candida auris*: a mini review on epidemiology in healthcare facilities in Asia. *J Fungi*. (2022) 8:1126. doi: 10.3390/jof8111126

21. Lee Y, Puumala E, Robbins N, Cowen LE. Antifungal drug resistance: molecular mechanisms in *Candida albicans* and beyond. *Chem Rev*. (2021) 121:3390–411. doi: 10.1021/acs.chemrev.0c00199

22. Garcia-Effron G. Molecular markers of antifungal resistance: potential uses in routine practice and future perspectives. *J Fungi*. (2021) 7:197. doi: 10.3390/jof7030197

23. Terra. Available at: https://app.terra.bio/ (Accessed March 23, 2023)

24. Voss K, Gentry J, Van der Auwera G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000Res*. (2017). doi: 10.7490/f1000research.1114631.1

25. cromwell: scientific workflow engine designed for simplicity & scalability. Trivially transition between one off use cases to massive scale production environments. Github. Available at: https://github.com/broadinstitute/cromwell (Accessed March 29, 2023)

26. miniwdl: workflow description language developer tools & local runner. Github. Available at: https://github.com/chanzuckerberg/miniwdl (Accessed March 29, 2023)

27. Bagal UR, Phan J, Welsh RM, Misas E, Wagner D, Gade L, et al. MycoSNP: a portable workflow for performing whole-genome sequencing analysis of *Candida auris* In: A Lorenz, editor. *Candida auris: Methods and protocols*. New York, NY: Springer US (2022). 215–28.

28. Seemann T. Nullarbor: "reads to report" for public health and clinical microbiology. Github. Available at: https://github.com/tseemann/nullarbor (Accessed March 29, 2023)

29. Gorzalski A, Ambrosio F, Massic L, Scribner M, Siao DD, Hau C, et al. The use of whole-genome sequencing and development of bioinformatics to monitor overlapping outbreaks of *C. auris* in southern Nevada. *Front Public Health*. (in press). doi: 10.3389/fpubh.2023.1198189

30. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. (2016) 17:132. doi: 10.1186/s13059-016-0997-x

31. Hall M. Rasusa: randomly subsample sequencing reads to a specified coverage. *J Open Source Softw*. (2022) 7:3941. doi: 10.21105/joss.03941

32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

33. BBMap. SourceForge (2022) Available at: https://sourceforge.net/projects/bbmap/ (Accessed March 23, 2023)

34. Seemann T. Shovill: ⚡♠ assemble bacterial isolate genomes from illumina paired-end reads. Github. Available at: https://github.com/tseemann/shovill (Accessed March 23, 2023)

35. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. (2012) 19:455–77. doi: 10.1089/cmb.2012.0021

36. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic *k*-mer extension for scrupulous assemblies. *Genome Biol*. (2018) 19:153. doi: 10.1186/s13059-018-1540-z

37. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. (2013) 29:1072–5. doi: 10.1093/bioinformatics/btt086

38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. (2015) 31:3210–2. doi: 10.1093/bioinformatics/btv351

39. Lumpe J, Gumbleton L, Gorzalski A, Libuit K, Varghese V, Lloyd T, et al. GAMBIT (Genomic Approximation Method for Bacterial Identification and Tracking): a methodology to rapidly leverage whole genome sequencing of bacterial isolates for clinical identification. *PLoS One*. (2023) 18:e0277575. doi: 10.1371/journal.pone.0277575

40. Seemann T. Snippy: rapid haploid variant calling and core genome alignment. Github. Available at: https://github.com/tseemann/snippy (Accessed March 23, 2023)

41. Li D, Wang Y, Hu W, Chen F, Zhao J, Chen X, et al. Application of machine learning classifier to *Candida auris* drug resistance analysis. *Front Cell Infect Microbiol*. (2021) 11:742062. doi: 10.3389/fcimb.2021.742062

42. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. (2016) 44:D733–45. doi: 10.1093/nar/gkv1189

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

44. Lumpe J. GAMBIT-publication: Snakemake workflow to generate figures and results from GAMBIT paper. Github. Available at: https://github.com/jlumpe/gambit-publication (Accessed March 23, 2023)

45. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. (2007) 9:90–5. doi: 10.1109/MCSE.2007.55

46. Yarmosh DA, Lopera JG, Puthuveetil NP, Combs PF, Reese AL, Tabron C, et al. Comparative analysis and data provenance for 1,113 bacterial genome assemblies. *mSphere*. (2022) 7:e0007722. doi: 10.1128/msphere.00077-22

47. Benton B, King S, Greenfield SR, Puthuveetil N, Reese AL, Duncan J, et al. The ATCC genome portal: microbial genome reference standards with data provenance. *Microbiol Resour Announc*. (2021) 10:e0081821. doi: 10.1128/MRA.00818-21

48. Home. ATCC Genome Portal Available at: https://genomes.atcc.org/?matchtype=&network=g&device=c&adposition=&keyword=&gclid=CjwKCAjw5pShBhB_EiwAvmnNV0VcXqyRYfC8xiNy1XCk_cpKPwqRHqOcxXJ2umhavjZhyN_wKkM2ixoCSV0QAvD_BwE (Accessed March 30, 2023)

49. Chow NA, Muñoz JF, Gade L, Berkow EL, Li X, Welsh RM, et al. Tracing the evolutionary history and global expansion of *Candida auris* using population genomic analyses. *mBio*. (2020) 11:e03364. doi: 10.1128/mBio.03364-19

50. Heath CH, Dyer JR, Pang S, Coombs GW, Gardam DJ. *Candida auris* sternal osteomyelitis in a man from Kenya visiting Australia, 2015. *Emerg Infect Dis*. (2019) 25:192–4. doi: 10.3201/eid2501.181321

51. Escandón P, Chow NA, Caceres DH, Gade L, Berkow EL, Armstrong P, et al. Molecular epidemiology of *Candida auris* in Colombia reveals a highly related, countrywide colonization with regional patterns in amphotericin B resistance. *Clin Infect Dis*. (2019) 68:15–21. doi: 10.1093/cid/ciy411

52. Hamprecht A, Barber AE, Mellinghoff SC, Thelen P, Walther G, Yu Y, et al. *Candida auris* in Germany and previous exposure to foreign healthcare. *Emerg Infect Dis*. (2019) 25:1763–5. doi: 10.3201/eid2509.190262

53. Rhodes J, Abdolrasouli A, Farrer RA, Cuomo CA, Aanensen DM, Armstrong-James D, et al. Genomic epidemiology of the UK outbreak of the emerging human fungal pathogen *Candida auris*. *Emerg Microbes Infect*. (2018) 7:43. doi: 10.1038/s41426-018-0045-x

54. Chow NA, de Groot T, Badali H, Abastabar M, Chiller TM, Meis JF. Potential fifth clade of *Candida auris*, Iran, 2018. *Emerg Infect Dis*. (2019) 25:1780–1. doi: 10.3201/eid2509.190686

55. Carolus H, Pierson S, Muñoz JF, Subotić A, Cruz RB, Cuomo CA, et al. Genome-wide analysis of experimentally evolved *Candida auris* reveals multiple novel mechanisms of multidrug resistance. *mBio*. (2021) 12:e03333. doi: 10.1128/mBio.03333-20

56. Tian S, Bing J, Chu Y, Chen J, Cheng S, Wang Q, et al. Genomic epidemiology of *Candida auris* in a general hospital in Shenyang, China: a three-year surveillance study. *Emerg Microbes Infect*. (2021) 10:1088–96. doi: 10.1080/22221751.2021.1934557

57. Burrack LS, Todd RT, Soisangwan N, Wiederhold NP, Selmecki A. Genomic diversity across *Candida auris* clinical isolates shapes rapid development of antifungal resistance in vitro and in vivo. *mBio*. (2022) 13:e0084222. doi: 10.1128/mbio.00842-22

58. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. (2019) 4:1686. doi: 10.21105/joss.01686

59. The R Project for Statistical Computing. Available at: https://www.r-project.org/ (Accessed June 26, 2023)

60. RStudio Team. *RStudio: integrated development for R RStudio, PBC*. Boston, MA: (2020) Available at: http://www.rstudio.com/.

61. Rybak JM, Sharma C, Doorley LA, Barker KS, Palmer GE, Rogers PD. Delineation of the direct contribution of *Candida auris* ERG11 mutations to clinical triazole resistance. *Microbiol Spectr*. (2021) 9:e0158521. doi: 10.1128/Spectrum.01585-21

62. Gold JAW, Ahmad FB, Cisewski JA, Rossen LM, Montero AJ, Benedict K, et al. Increased deaths from fungal infections during the coronavirus disease 2019 pandemic-National Vital Statistics System, United States, January 2020-December 2021. *Clin Infect Dis*. (2023) 76:e255–62. doi: 10.1093/cid/ciac489

63. Leber AL, Peterson E, Dien Bard J. Personnel standards and workforce subcommittee, American Society for Microbiology. The hidden crisis in the times of COVID-19: critical shortages of medical laboratory professionals in clinical microbiology. *J Clin Microbiol*. (2022) 60:e0024122. doi: 10.1128/jcm.00241-22

64. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*. (2015) 31:2877–8. doi: 10.1093/bioinformatics/btv271

65. Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, den Bakker HC, et al. Mashtree: a rapid comparison of whole genome sequence files. *J Open Source Softw*. (2019) 4:10.21105/joss.01762. doi: 10.21105/joss.01762

66. Chakrabarti A, Sood P. On the emergence, spread and resistance of *Candida auris*: host, pathogen and environmental tipping points. *J Med Microbiol*. (2021) 70:001318. doi: 10.1099/jmm.0.001318

67. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat Med*. (2020) 26:832–41. doi: 10.1038/s41591-020-0935-z

frontiers | Frontiers in Public Health

TYPE Perspective
PUBLISHED 03 August 2023
DOI 10.3389/fpubh.2023.1204064

EDITED BY
Peera Hemarajata,
Association of Public Health Laboratories,
United States

REVIEWED BY
Varvara K. Kozyreva,
California Department of Public Health,
United States
Angela Starks,
Centers for Disease Control and Prevention
(CDC), United States

*CORRESPONDENCE
Daniela Maria Cirillo
✉ cirillo.daniela@hsr.it

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

‡These authors share last authorship

RECEIVED 11 April 2023
ACCEPTED 13 July 2023
PUBLISHED 03 August 2023

CITATION
de Araujo L, Cabibbe AM, Mhuulu L, Ruswa N, Dreyer V, Diergaardt A, Günther G, Claassens M, Gerlach C, Utpatel C, Cirillo DM, Nepolo E and Niemann S (2023) Implementation of targeted next-generation sequencing for the diagnosis of drug-resistant tuberculosis in low-resource settings: a programmatic model, challenges, and initial outcomes.
Front. Public Health 11:1204064.
doi: 10.3389/fpubh.2023.1204064

Implementation of targeted next-generation sequencing for the diagnosis of drug-resistant tuberculosis in low-resource settings: a programmatic model, challenges, and initial outcomes

Leonardo de Araujo[1†], Andrea Maurizio Cabibbe[2†], Lusia Mhuulu[3†], Nunurai Ruswa[4], Viola Dreyer[1], Azaria Diergaardt[3], Gunar Günther[3,5], Mareli Claassens[3], Christiane Gerlach[1], Christian Utpatel[1], Daniela Maria Cirillo[2*‡], Emmanuel Nepolo[3‡] and Stefan Niemann[1,3‡]

[1]Molecular and Experimental Mycobacteriology Group, Research Center Borstel, Leibniz Lung Center, Borstel, Germany, [2]Emerging Bacterial Pathogens Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy, [3]Department of Human, Biological & Translational Sciences, School of Medicine, University of Namibia, Windhoek, Namibia, [4]National TB and Leprosy Programme, Ministry of Health and Social Services, Windhoek, Namibia, [5]Department of Pulmonology and Allergology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Targeted next-generation sequencing (tNGS) from clinical specimens has the potential to become a comprehensive tool for routine drug-resistance (DR) prediction of *Mycobacterium tuberculosis* complex strains (MTBC), the causative agent of tuberculosis (TB). However, TB mainly affects low- and middle-income countries, in which the implementation of new technologies have specific needs and challenges. We propose a model for programmatic implementation of tNGS in settings with no or low previous sequencing capacity/experience. We highlight the major challenges and considerations for a successful implementation. This model has been applied to build NGS capacity in Namibia, an upper middle-income country located in Southern Africa and suffering from a high-burden of TB and TB-HIV, and we describe herein the outcomes of this process.

KEYWORDS

Next generation sequencing (NGS), *Mycobacterium tuberculosis* complex (MTBC), NGS clinical use, Genomic diagnostic, Genomic surveillance, NGS programmatic implementation, high TB burden countries, low- and middle-income countries

1. Introduction

Infectious diseases are currently one of the most explored fields for clinical and public health genomics, as sequencing technologies simplified and accelerated the deep characterization of pathogens (1). Pathogen genomics is transforming surveillance programs allowing both prompt identification of outbreaks and epidemics and accurate diagnosis at individual level, replacing the standard techniques in microbiology laboratories (2, 3). The emergence of infectious threats, such as SARS-CoV-2 and Monkeypox viruses, showed the needs of strengthening health systems worldwide with implementation of NGS capacity (4). However, other more prevalent diseases such as tuberculosis (TB) and malaria should also profit on NGS implementation to improve diagnosis and for monitoring/surveillance.

TB is a leading infectious killer, after COVID-19 in 2020/2021, with estimated total incidence of 10.6 million new cases and 1.6 million deaths (5). It is also the leading killer of people living with human immunodeficiency virus (HIV) and a major contributor to deaths related to antimicrobial resistance. The use of World Health Organization (WHO)-recommended molecular diagnostics (mWRDs) for diagnosis and drug resistance (DR) testing to at least key drugs such as rifampicin, isoniazid and fluoroquinolones, remains limited in low-resource, high TB burden settings. The incomplete drug sensitivity testing (DST) coverage leads to empiric treatment initiation, despite the existing treatment guidelines requiring access to testing (6).

Whilst mWRDs already accessible in low- and middle- income countries (LMICs) allow prompt resistance prediction for one or few drugs, next generation sequencing (NGS) of *Mycobacterium tuberculosis* complex (MTBC) strains offers the most comprehensive approaches to determine resistance to the current recommended regimens (3, 6). Two main NGS-based approaches may be used: whole genome sequencing (WGS) and targeted NGS (tNGS). tNGS takes advantage of the selective amplification of DR-related regions of MTBC genome and provides quick results directly from clinical specimen, with higher sensitivity than WGS, lower turnaround time and easier interpretation (7–10).

Genome sequencing has also already been introduced as a tool to investigate TB DR evolution, transmission dynamics and the population structure of MTBC, for surveillance of DR (9–11), and patient's management (12), although a roadmap to a programmatic implementation of TB genomics is still lacking.

Recent investigations have shown that using sequencing to inform treatment regimens for DR TB led to decisions comparable to those derived from phenotypic DST (pDST) (6, 7). Also, it became evident that analyzing by NGS all genes known to be associated with DR would improve the design of personalized multidrug-resistant (MDR) TB regimens (high concordance with pDST-informed decisions) (13). However, the implementation of NGS in TB clinical laboratories requires adequate infrastructure, training, and strategic planning. Challenges include procurement, sample referral, quality-assured procedures, data management, translation into clinical practice and sustainability (e.g., human resource retention) (14, 15). Therefore, it is important to collect data-driven evidence from practical implementations in high TB burden countries. Many of the challenges to an effective implementation of genomics in resource-limited settings are technical and deal with the renovation of healthcare systems, including high costs, suboptimal supply, inadequate infrastructure and link of sequencing information to existing record systems (1, 16, 17). Other aspects involve social and ethical components (use/sharing of data and clinical application thereof).

Herein we detail our model of implementation of tNGS for DR prediction of MTBC strains in settings with no or low previous sequencing capacity. We detail how this model was implemented at the University of Namibia (UNAM) in Windhoek, Namibia, one of the 30 high-burden countries for TB and TB-HIV.

2. Implementation model

In this section we define the steps that we consider crucial and how those were addressed during the tNGS implementation in Namibia (*in italics*).

2.1. Implementation strategy and roadmap

As shown in Figure 1A, our strategy for the implementation of NGS was based on three pillar phases: preparation, implementation, and sustainability. These phases are subdivided into smaller categories of tasks (Figure 1A, white boxes).

In the preparation phase, as the first phase of the implementation process, we defined the main outcomes of the intervention and the strategy to assess deliverables. The entire outline of the implementation process must be planned here.

The implementation phase focuses on practical work once the strategies have been developed. Capacity building, support, training, pilots and the search for sustainability begin in this phase.

The sustainability phase aims to scale up NGS capacity, anchor NGS in local guidelines and help programs in the search for new sources of funding.

Our implementation strategy was based on the exchange of knowledge between a center of expertise for NGS (non-profitable), in this case with extensive experience in doing NGS on clinical MTBC strains, to another center in a LMIC that does not had this capacity.

2.1.1. Preparation phase

2.1.1.1. Site identification

Epidemiological context, laboratory network and testing algorithms are considered in our approach. In this context, we needed to consider the local testing algorithm for TB and DR TB, in order to determine the best way to incorporate tNGS. We needed to identify which specimens could be used to extract mycobacterial DNA (e.g., leftovers of sputum specimens, new sputum, cultures, etc.) by avoiding unnecessary additional steps to the standard procedures for collecting and preparing specimens.

Namibia has an estimated TB incidence of 460/100,000 population and estimated 560 multidrug-resistant (MDR) TB cases per year. Xpert MTB/RIF Ultra and line probe assays (LPA) 1st-2nd line are used. The diagnostic algorithm is reported in Supplementary Text*. Second line DST testing coverage is incomplete due to reagent stock outs or culture contamination leading to shipment of selected strains for further testing outside the country. The NGS platform was implemented at an academic institution, the UNAM, within the national TB programme (NTP) network. Consultative need assessment meetings were held between the UNAM, the National TB and Leprosy Programme and collaborating stakeholders (Research Center Borstel and Robert Koch Institute, Germany, Ministry of Health and Social Services and the Namibian Institute of Pathology) which facilitated and supported the implementation of tNGS in Namibia. Additionally, in Namibia the use*

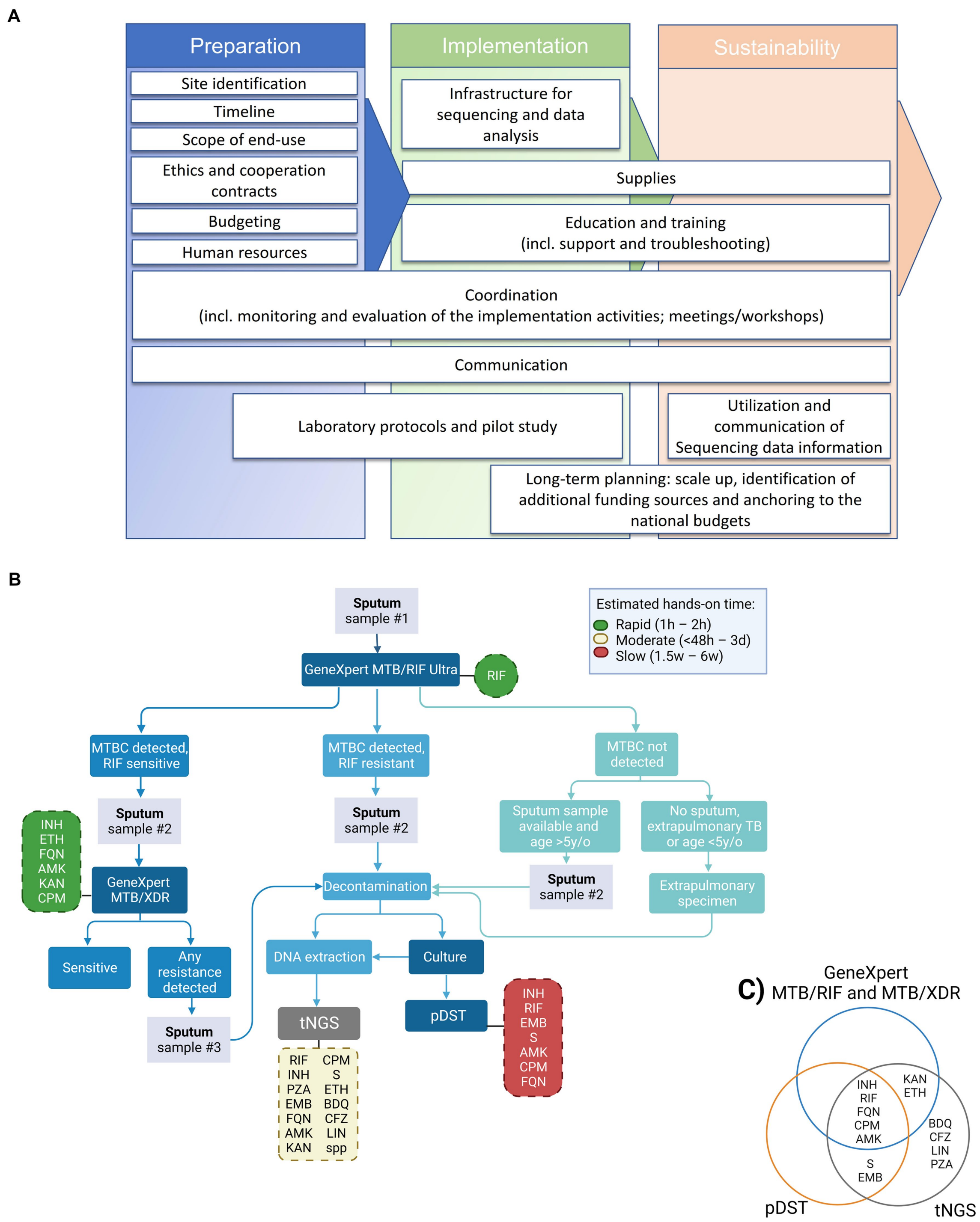

A
Preparation
Implementation
Sustainability
Site identification
Timeline
Scope of end-use
Ethics and cooperation contracts
Budgeting
Human resources
Infrastructure for sequencing and data analysis
Supplies
Education and training (incl. support and troubleshooting)
Coordination (incl. monitoring and evaluation of the implementation activities; meetings/workshops)
Communication
Laboratory protocols and pilot study
Utilization and communication of Sequencing data information
Long-term planning: scale up, identification of additional funding sources and anchoring to the national budgets
B
Sputum sample #1
GeneXpert MTB/RIF Ultra
RIF
Estimated hands-on time:
🟩 Rapid (1h – 2h)
🟨 Moderate (<48h – 3d)
🟥 Slow (1.5w – 6w)
MTBC detected, RIF sensitive
MTBC detected, RIF resistant
MTBC not detected
Sputum sample #2
INH ETH FQN AMK KAN CPM
GeneXpert MTB/XDR
Sputum sample #2
Sputum sample available and age >5y/o
No sputum, extrapulmonary TB or age <5y/o
Sputum sample #2
Extrapulmonary specimen
Decontamination
Sensitive
Any resistance detected
DNA extraction
Culture
Sputum sample #3
tNGS
pDST
RIF INH PZA EMB FQN AMK KAN CPM S ETH BDQ CFZ LIN spp
INH RIF EMB S AMK CPM FQN
C) GeneXpert MTB/RIF and MTB/XDR
INH RIF FQN CPM AMK
KAN ETH
BDQ CFZ LIN PZA
S EMB
pDST
tNGS

FIGURE 1
Proposed tNGS implementation roadmap, diagnostic flowchart and tested drugs for the detection of drug-resistance TB cases including NGS technologies for enhanced genotypic resistance prediction of *M. tuberculosis* complex (MTBC) strains. **(A)** Roadmap for building tNGS capacity in low- and middle-income countries (from left to right). **(B)** Flowchart including the GeneXpert MTB/RIF Ultra and MTB/XDR (or any other endorsed test with similar application) as screening tests for selection of samples to downstream targeted next-generation sequencing (tNGS). Hands-on time refers specifically to the theoretical amount of time needed to perform the test, excluding the time needed to collect the clinical specimen. Turn-around time (TAT) refers to the theoretical amount of time needed to perform the test added to the theoretical time to have available clinical specimens (specially affected when additional samples have to be collected). The TAT for tNGS in routine settings is expected to be around 6–8 days from the entry test result. Dashed boxes indicate the panel of drug-resistances investigated by each test, green, yellow and red colors indicate the relative hands-on time, reference: (12). **(C)** Venn diagram depicting the panel of drug-resistances investigated by each test, worth mentioning that the number of amplified targets is different between the molecular DST tests. pDST, phenotypic drug sensitivity testing; INH, isoniazid; RIF, rifampicin; FQN, Fluoroquinolones; CPM, capreomycin; AMK, amikacin; S, streptomycin; EMB, ethambutol; KAN, kanamycin; ETH, ethionamide; BDQ, bedaquiline; CFZ, clofazimine; LIN, linezolid; PZA, pyrazinamide.

of tNGS for TB diagnosis is supported by the NTP as one recommendation of the end term review of the NTP is "to develop a contingency plan and diversify TB testing, particularly drug resistance testing."

2.1.1.2. Timeline

The complete timeline must be realistic and focused on priorities to mark the overall project progress.

Our NGS implementation process in Namibia started in 2019 with a 1-year long preparation phase, during this period, among other tasks, the protocols were developed and tested at the center of expertise for NGS. After that, the capacity building started, i.e., the implementation phase, with an initial aim to be done in 2-years. However, our timeline was severely affected by COVID-19 pandemic, the implementation phase was duly extended having a total duration of 4 years (2019–2022).

2.1.1.3. Scope of end-use

The overarching goal of the implementation process is to develop local technical capacity for the use of tNGS as an add-on diagnostic tool for prediction of DR profiles of medicines included in MDR or rifampicin-resistant (RR) TB regimens. A proposed workflow for the future incorporation of tNGS into the national algorithm is shown on Figure 1B. Our proposed algorithm would include samples for tNGS that presented prior DR profile on screening test(s), i.e., GeneXpert MTB/RIF Ultra and MTB/XDR. tNGS would enable faster results than culture-based pDST, with the additional benefit of interrogating genotypic resistance to bedaquiline, clofazimine, linezolid and pyrazinamide (currently not tested in the local settings, Figure 1C).

Initially, around 50 MDR/RR TB samples are expected per year mainly from the TB referral hospital in Windhoek and will be used for pilot adoption of tNGS, then it is planned to roll out to cover MDR/RR-TB cases from the entire country.

2.1.1.4. Ethics and cooperation contracts

Ethical review ensures that the study adheres to the agreed ethical standards and in this field with the specific consideration of sharing of genetic data and personal data.

The project in Namibia was approved by the local Ethics committee of UNAM and approved by the Ministry of Health (# 17/3/3EN). Cooperation and material transfer agreements were also signed by the implementing partners and the implementation site, containing the scope of the process and terms for collaboration.

2.1.1.5. Budgeting

A financial plan must include costs for key actors and activities including personnel, infrastructure, purchase of devices and consumables, shipments, training (including travel and accommodation), internet and other services.

In Supplementary Figure 1 we reported a rough average percentage of expenses per year over the 4 years of implementation in Namibia. Human resources (HR) at the referral center and implementation sites are the highest expense (52%), followed by consumables (23%), and devices (17%). The total funds used in this period was approximately 418,500.00 USD. With regards to equipment maintenance, as we acquired new equipment, it came with the standard manufacturer's warranty for at least the initial implementation phase. After that, extended warranty plans were considered for just the sequencing devices and were contracted with the official local distributors of the sequencers. The other equipment was included in the regular institutional maintenance activities, being under UNAM responsibility.

2.1.1.6. HR

A survey of available laboratory staff must be carried out to understand if reallocation of existing staff is feasible. If new staff is hired, training needs should be considered. The calculation of needed personnel was done empirically based on our experience as center of expertise for NGS, available funds, and expected sample flow.

A laboratory technologist and a PhD student were recruited to be dedicated to the development of tNGS activities locally, based on their background in medical laboratory sciences and medical biochemistry, with molecular biology techniques experience such as Sanger sequencing.

2.1.1.7. Coordination

Defining the coordination process is an essential part, coordinators supervise all processes (HR, materials and equipment) and ensure that all team members are aware of the objectives, schedule and progress. Resources must be allocated efficiently, potential risks anticipated and mitigation strategies applied.

Implementation coordinators were hired/delegated at the international center of expertise for NGS and locally.

2.1.1.8. Communication

Effective communication of results is a key step to motivate all stakeholders (Ministry of Health (MoH), staff and the community).

Annual workshops were carried to share the implementation outcomes with main stakeholders. Concurrently the tNGS was discussed regularly with NTP and clinical partners in order to secure an early translation of implementation into clinical practice (ongoing), based on the tNGS data currently being generated in-country during implementation.

2.1.2. Implementation phase

2.1.2.1. Infrastructure for sequencing and data analysis

This capacity involves the selection of the adequate facilities for the installation of the NGS lab. A wet lab infrastructure herein refers to the physical laboratory space, equipment, and reagents required for the pre-PCR area (DNA extraction), and the post-PCR area (PCR amplification, library preparation, and sequencing). The sequencing laboratory should have adequate space, sturdy benchtops, electricity and internet outlets, and follow strict room temperature, humidity, and air quality requirements for operation of the sequencers and related instruments.

The sequencing apparatus was installed in the lab of the department of Human, Biological and Translational Medical Sciences in collaboration with the "Group Research in Infectious Diseases (GRID), UNAM (Supplementary Figure 2). Two Illumina iSeq100 machines were purchased as an upgrade to the existing Sanger sequencing facility. The iSeq100 instrument was selected based on its size, multiplexing capacity suitable for local workflows, cost-effectiveness, and user-friendly *interface. The user's manual was referred to find the appropriate site for installation.*

A dry lab infrastructure refers to the computational infrastructure required for analyzing and interpreting the sequencing data generated by tNGS.

*The sequencers were connected to the network, to a network attached storage system for data backup and to an uninterrupted power supply device. Fridge and freezers for the reagent's storage and a thermal cycler were also procured. Due to power interruption experienced at the implementation site a backup freezer was installed with backup power supply. Computers were purchased for routine use, which are sufficient for analyzing tNGS in the cloud. Additionally, a high-end computer was purchased to handle more complex local analysis in case further needs are identified (*Supplementary Table 1*).*

2.1.2.2. Supplies

An efficient and trustworthy supply chain is crucial for achieving sustainability.

Materials were securely delivered either from the Research Center Borstel, Germany, and shipped, or through regional distributors (however mainly located in South Africa).

*We performed a rough estimation of the initial investment needed to procure the consumables required to start the tNGS activities (*Supplementary Table 2*).*

2.1.2.3. Education, training, and support

After the training and competency assessment, continued assistance should be conducted (using instant messaging apps, regular meetings, and written reporting methods).

In Namibia, personnel were trained on tNGS workflow and analysis, including hands-on, quality checks, and run of the iSeq100 instrument. Training was performed either hybrid or locally at the implementing facility. The educational activities comprehensively included theoretical and practical packages with tailored agenda, refresh, and troubleshooting sessions over the entire implementation phase. Data analysis requested additional tutoring activities and was not limited to the laboratory study staff.

2.1.2.4. Laboratory protocols and pilot study

Upon completion of the practical sequencing training, local samples from TB patients are sequenced as pilot to assess the capacity and feasibility.

Strains from clinical culture TB samples were subjected to tNGS as pilot (No. 48 RR as identified by Xpert MTB/RIF Ultra). The details and outcomes of this study are described in Supplementary Text *and* Supplementary Table 3.

3. Discussion

NGS-based analysis of clinical MTBC strains bridge gaps associated with pDST and the limitations of other mWRDs for DR testing (12). tNGS can detect DR-associated mutations to a variety of antibiotics as those included in the WHO-recommended DR TB treatments, and is applied to a variety of clinical specimens (9, 10, 18). It also allows species identification at the lineage level, detection of mixed populations and heteroresistance (9, 10, 18).

It is expected that within the next few years, with an increased automation of the NGS workflow and improved treatment algorithms, NGS workflows will be widely implemented at least at reference laboratory level and clinicians will adopt individualized treatment decisions soon after diagnosis of DR TB is made by entry tests from patient samples (ideally 6–8 days in programmatic conditions), adjusting also the duration of treatments, and therefore their efficacy, costs and toxicity (6).

The process of implementation of TB genomics for surveillance of DR TB, recommended by the WHO, and as a routine diagnostic tool at country level, requires careful planning, strong commitment, and investment to support successful adoption into national algorithms. Before the implementation process starts, it is crucial to ponder the use of NGS within the NTP, with implications for choice of technologies and equipment to use, selection of sites, referral systems, target turnaround times, implementation of clinical decision making and incorporation into treatment guidelines. An implementation plan is needed to build the NGS infrastructures (wet and dry) and HR (management and technical). In addition to a careful planning step, it is also important to proceed stepwise and to report the obstacles identified during the implementation process. If the implementation strategy is not adequate, based on evidence, and without regular assessments, this tool will not positively impact the diagnosis at implementation sites. In fact, considering the several barriers to NGS uptake at country level, a completely 'self-sustainable' NGS capacity seems still far from reality in LMICs (15). Reassuringly, laboratory infrastructures and specialized academic education for scientists and clinicians are expanding quickly in LMICs, opening new opportunities for research in such scenarios, further progress, awareness and equitable partnerships (19, 20). With the COVID-19 pandemic, the number of countries recognized by the African Union having local access to sequencing facilities increased by around 50% and consortia were created to study the spread of SARS-CoV-2 variants (21). Worth mentioning that the NGS capacity developed within the project in Namibia was timely used for emergency response to COVID-19 pandemic and variant's surveillance. In this context and given costs, countries should consider that an investment for a given disease (e.g., surveillance of DR TB or COVID-19) can impact other health priorities such as viral diseases or AMR surveillance, as it creates facilities and capacity that can be expanded. Another unforeseen benefit was that the trained personnel at UNAM were able to train other laboratory scientists from the Sub-Saharan region (Botswana, and Eswatini).

Some challenges to a programmatic implementation of NGS are technical and include the complexity of protocols and workflows, and sophisticated and not yet fully standardized data management, analysis and interpretation. To tackle this, several collaborative initiatives were created in order to provide data on the performance of existing technologies, and industries and researchers are developing end-to-end user-friendly NGS solutions. Furthermore, knowledge of molecular mechanisms at the basis of the emergence of DR is incomplete, limiting the predictive values of NGS. The correct use of data requires training of clinicians on their interpretation (Figure 1A, utilization and communication of sequencing data information), a competence that in initiated but still faces challenges in Namibia. Herein, the analysis of our pilot data indicates that the protocols need to be properly validated at the implementation sites, and that the limitations of tNGS have to be evaluated locally. We had to start the implementation with DNA from MTBC cultures for ease, although the immediate next step is to apply it on primary patient samples (e.g., sputum) with protocols already validated. On the brighter side, the analysis of the pilot samples showed a clear advantage of tNGS by providing a broader vision of resistance profile to anti-TB drugs (Figure 1C), the user-friendly analysis interface, validation steps and the indication of the "usability" of the sequencing results.

The main challenges of the NGS implementation process at UNAM were: (i) initial technical issues experienced with the iSeq100 setup, later resolved through remote technical support; (ii) unforeseen costs to stabilize and control the room temperature required due to the local semi-arid environment; (iii) an unstable internet connection has resulted in a challenging data upload process; (iv) delay in the construction of dedicated pre-PCR and post-PCR areas; (v) delay in the delivery of material due to COVID-19 restrictions; and (vi) sufficient adhesion of local stakeholders to translate tNGS results into clinical practice and public health policy. Despite these challenges, our planning and initial implementation phase are finalized, NGS capacity was successfully built and is currently in use by the GRID, UNAM, while the implementation of tNGS results into clinical processes, and some other competencies of the sustainability phase, have just started. We recognize the need for a clinical advisory committee (CAC), that will review and discuss the reports generated by tNGS, as well as provide guidance to the clinicians. The CAC should consist of representatives from NTP and National TB Reference Laboratories, implementing partners, laboratory specialists, clinicians, and international experts. However, the implementation of such committees and the approval to use tNGS data requires multisectoral and political support in the country. This process is ongoing and challenging, but it should be facilitated upon the release of WHO guidelines for tNGS use in DR TB diagnostics. The plan is that the tNGS DR report and standard results will be shared with the CAC, which will review and discuss the data and provide guidance to the clinicians.

Logistic aspects, such as procurement and supply chains in countries where distributors are not present and unable to provide optimal maintenance and support, importation requirements, as well as transportation of samples or reagents in case of unreliable referral systems, represent threats in the current scenario. In Namibia, the obstacles primarily revolve around the market and logistic, there are complete shortages of some products in the local market. As a solution, we have managed to identify suppliers of sequencing products in neighboring countries, particularly South Africa. However, these are only third-part distributors, and the materials are imported from other countries. Consequently, there is a significant loss of shelf-life during transportation between manufacturer-distributor- implementation site; increased prices are also expected due to the reselling process. As a feasible alternative, but not long-term sustainable, the products can be purchased at the center of expertise for NGS outside the country (in a place where the supplier availability, market prices and logistics are less challenging), re-packed and sent to the implementation site. Another important, yet often unforeseen obstacle that greatly affects the availability of NGS supplies, is the bureaucratic and time-consuming processes for importation of donated goods. This still requires facilitation by the local authorities. WHO and other TB stakeholders are encouraging companies involved in the manufacturing/supply of NGS-related items to explore ways to expand the use of genomics in LMICs and make it more accessible, through incentives such as modified pricing models, reduction of costs and loans at low-interest (22). The development of reagents/technologies that do not require temperature control can also help to push the implementation of NGS in Sub-Saharan Africa.

TABLE 1 SWOT analysis for use of tNGS-based genomics in TB control and care.

Strengths	Weaknesses
❖ Multi-purpose, multi-disease	❖ Need of genotypic-phenotypic associations
❖ Suitable from various sample types	❖ Turnaround time depends on sample referral and sequencing capacity/multiplexing
❖ Rapid (faster turnaround times than conventional pDST testing)	❖ Start-up costs
❖ Kit-based and user-friendly analysis tools (improved standardization)	❖ Currently not feasible at peripheral level
❖ Deep level of genetic information enabling "precision"	❖ Procurement and supply chain
	❖ Need of specialized and trained personnel
Opportunities	**Threats**
❖ Less phenotyping in routine testing	❖ Borderline mutations
❖ High predictive value for drug-resistance	❖ Confidence-grading of mutations requires large and representative datasets
❖ Huge research on innovative NGS technologies	❖ Support to clinicians
❖ Development of lists of confidence-graded mutations reflecting on routine Nucleic Acid Amplification Tests	❖ Not all resistance mechanisms can be explored (e.g., gene expression, structural changes)
❖ Interrogates resistance to additional anti-microbials not routinely tested in national algorithms	❖ Information technology (IT) infrastructure
❖ Research outcomes	❖ Cost-effectiveness to be demonstrated
	❖ Efficient and timely results reporting
	❖ To achieve sustainability

This table provides a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis of the use of genomics in TB control and care. The SWOT analysis helps to identify strategies to maximize the advantages and mitigate the disadvantages of having tNGS capacity for TB implemented.

The SWOT (strengths, weaknesses, opportunities, and threats) analysis reported in Table 1 offers a framework to assess the characteristics of the TB genomics solution and helps to identify the main areas for improvement, and strategies to maximize the advantages and to mitigate the disadvantages of tNGS in TB control and care. We proposed here a model based on tNGS implementation, which is more suitable than WGS for direct and faster DST testing, as it usually does not require culture. This approach offers higher standardization of wet protocols as kit-based, automated data analysis pipeline, and increased multiplexing compared to WGS for routine settings. Conversely, WGS approach from cultures would provide higher resolution of genomes and transmission outbreaks, but currently with a more challenging implementation process in high TB burden, low-resource settings.

The ultimate scope of the tNGS implementation in routine settings of high TB and/or DR TB incidence countries is to improve clinical management of cases and provide surveillance to resistance to new regimens. Relevant TB stakeholders are looking with interest at the pilot implementation studies and findings and should encourage roll out at the level of national/regional laboratories (23). It is advisable that, once tNGS receives approval from WHO as a diagnostic tool for DR TB, NTPs will include tNGS in the diagnostic algorithms. This should be accompanied by sustainability plans and budgetary allocations. In settings where NGS capacity is lacking, the implementation process can be better planned by leveraging the experience presented in this study.

Sustainability can be achieved by several measures, starting with the release of WHO policies on NGS use for clinical care/surveillance, then with the inclusion in the Global Drug Facility list for regular global supply at negotiated price, the adoption of NGS into national guidelines, and the development of more cost-effective protocols and other commercial point-of-care solutions. Furthermore, the incorporation of sequencing as valuable public health tool into the national anti-TB DR surveys will facilitate the collection of comprehensive data from countries. This data will inform timely public health actions to be integrated into national strategic plans for TB. It will enable the design of optimized diagnostic algorithms, assessment of the efficacy of recommended treatment regimens, identification of research needs, and guide resource allocation planning (24).

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 at: https://www.ebi.ac.uk/ena/, PRJEB62858.

Author contributions

LA, AC, LM, NR, VD, GG, CG, DC, EN, and SN conceived the idea and designed the study. LA, AC, LM, VD, CU, EN, and SN analyzed and interpreted the data. LM, AD, VD, and CU generated the sequencing data. LA, LM, AD, GG, MC, CG, EN, and SN coordinated the implementation. LA, LM, AC, GG, and NR designed the proposed diagnostic workflows. LA, AC, and LM wrote the initial draft of the manuscript. DC, EN, and SN supervised the study. All authors contributed to obtaining and assembling the data, during the review process, data interpretation, critical review of the manuscript, and approved the final version of the manuscript.

Funding

Implementation of NGS in the country was supported by the German Global Health Protection Program – Federal Ministry of Health, the SeqMDRTB_NET (ZMVI1-2519GHP708, https://ghpp.de/en/projects/seqmdrtb-net/).

Acknowledgments

The authors are grateful to all the scientific and management staff from the University of Namibia and Research Center Borstel, Leibniz Lung Center that contributed to the success of this project.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1204064/full#supplementary-material

References

1. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, et al. From public health genomics to precision public health: a 20-year journey. *Genet Med*. (2018) 20:574–82. doi: 10.1038/gim.2017.211

2. Juan Carlos G-V, Nadia Alejandra R-S. *Principles of genetics and molecular epidemiology*. Cham: Springer Nature (2022).

3. Domínguez J, Boeree MJ, Cambau E, Chesov D, Conradie F, Cox V, et al. Clinical implications of molecular drug resistance testing for *Mycobacterium tuberculosis*: a 2023 Tbnet/RESIST-TB consensus statement. *Lancet Infect Dis*. (2023) 23:E122–37. doi: 10.1016/S1473-3099(22)00875-1

4. Jaiswal V, Nain P, Mukherjee D, Joshi A, Savaliya M, Ishak A, et al. Symptomatology, prognosis, and clinical findings of monkeypox infected patients during COVID-19 era: a systematic-review. *Immun Inflam Dis*. (2022):e722:10. doi: 10.1002/iid3.722

5. WHO. *Global tuberculosis report 2022 N.D*. Geneva: WHO. (2023).

6. Lange C, Alghamdi WA, Al-Shaer MH, Brighenti S, Diacon AH, Dinardo AR, et al. Perspectives for personalized therapy for patients with multidrug-resistant tuberculosis. *J Intern Med*. (2018) 284:163–88. doi: 10.1111/joim.12780

7. Gröschel MI, Walker TM, Van Der Werf TS, Lange C, Niemann S, Merker M. Pathogen-based precision medicine for drug-resistant tuberculosis. *PLoS Pathog*. (2018) 14:E1007297. doi: 10.1371/journal.ppat.1007297

8. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol*. (2019) 17:533–45. doi: 10.1038/S41579-019-0214-5

9. Jouet A, Gaudin C, Badalato N, Allix-Béguec C, Duthoy S, Ferré A, et al. Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to 13 anti-tuberculous drugs. *Eur Respir J*. (2021) 57:2002338. doi: 10.1183/13993003.02338-2020

10. Feuerriegel S, Kohl TA, Utpatel C, Andres S, Maurer FP, Heyckendorf J, et al. Rapid genomic first- and second-line drug resistance prediction from clinical *Mycobacterium tuberculosis* specimens using DEEPLEX-Myctb. *Eur Respir J*. (2021) 57:2001796. doi: 10.1183/13993003.01796-2020

11. WHO. *Guidance for the surveillance of drug resistance in tuberculosis*. Geneva: WHO (2020).

12. Dookie N, Khan A, Padayatchi N, Naidoo K. Application of next generation sequencing for diagnosis and clinical management of drug-resistant tuberculosis:

updates on recent developments in the field. *Front Microbiol*. (2022) 13:775030. doi: 10.3389/fmicb.2022.775030

13. Finci I, Albertini A, Merker M, Andres S, Bablishvili N, Barilar I, et al. Investigating resistance in clinical *Mycobacterium tuberculosis* complex isolates with genomic and phenotypic antimicrobial susceptibility testing: a multicentre observational study. *Lancet Microbe*. (2022) 3:E672–82. doi: 10.1016/S2666-5247(22)00116-1

14. Cabibbe AM, Walker TM, Niemann S, Cirillo DM. Whole genome sequencing of *Mycobacterium tuberculosis*. *Eur Respir J*. (2018) 52:1801163. doi: 10.1183/13993003.01163-2018

15. Vogel M, Utpatel C, Corbett C, Kohl TA, Iskakova A, Ahmedov S, et al. Implementation of whole genome sequencing for tuberculosis diagnostics in a low-middle income. *High MDR-TB Burden Country Sci Rep*. (2021) 11:15333. doi: 10.1038/S41598-021-94297-Z

16. Flahault A, Utzinger J, Eckerle I, Sheath DJ, De Castañeda RR, Bolon I, et al. Precision global health for real-time action. *Lancet Digital Health*. (2020) 2:E58–9. doi: 10.1016/S2589-7500(19)30240-7

17. Ho D, Quake SR, Mccabe ERB, Chng WJ, Chow EK, Ding X, et al. Enabling technologies for personalized and precision medicine. *Trends Biotechnol*. (2020) 38:497–518. doi: 10.1016/j.tibtech.2019.12.021

18. Sibandze DB, Kay A, Dreyer V, Sikhondze W, Dlamini Q, Dinardo A, et al. Rapid molecular diagnostics of tuberculosis resistance by targeted stool sequencing. *Genome Med*. (2022) 14:52. doi: 10.1186/S13073-022-01054-6

19. Temesgen Z, Cirillo DM, Raviglione MC. Precision medicine and public health interventions: tuberculosis as a model? *Lancet Public Health*. (2019) 4:E374. doi: 10.1016/S2468-2667(19)30130-6

20. Ntoumi F, Petersen E, Mwaba P, Aklillu E, Mfinanga S, Yeboah-Manu D, et al. Blue skies research is essential for ending the tuberculosis pandemic and advancing a personalized medicine approach for holistic management of respiratory tract infections. *Int J Infect Dis*. (2022) 124:S69–74. doi: 10.1016/j.ijid.2022.03.012

21. Kwon D. 100,000 coronavirus genomes reveal COVID's evolution in Africa. *Nature*. (2022). doi: 10.1038/d41586-022-03070-3, [Epub ahead of print]

22. WHO. *WHO science council meeting, Geneva, Switzerland, 11–12 July 2022: report*. Geneva: World Health Organization (2022).

23. High-tech tests for drug-resistant TB [Internet]. Unitaid. (2019) [cited 2023 Jul 17]. Available from: https://unitaid.org/project/hight-tech-tests-for-drugresistent-tb/#en

24. Dean AS, Tosas Auguet O, Glaziou P, Zignol M, Ismail N, Kasaeva T, et al. 25 years of surveillance of drug-resistant tuberculosis: achievements, challenges, and way forward. *Lancet Infect Dis*. (2022) 22:E191–6. doi: 10.1016/S1473-3099(21)00808-2

25. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*. (2015) 10:E0128036. doi: 10.1371/journal.pone.0128036

26. Pinnock H, Barwick M, Carpenter CR, Eldridge S, Grandes G, Griffiths CJ, et al. Standards for reporting implementation studies (Stari) statement. *BMJ*. (2017):i6795. doi: 10.1136/bmj.i6795

27. Van Soolingen D, Hermans PW, De Haas PE, Soll DR, Van Embden JD. Occurrence and stability of insertion sequences in *Mycobacterium Tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol*. (1991) 29:2578–86. doi: 10.1128/jcm.29.11.2578-2586.1991

28. Crawford JT, Eisenach KD, Bates JH. Diagnosis of tuberculosis: present and future. *Semin Respir Infect*. (1989) 4:171–81.

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 23 October 2023
DOI 10.3389/fpubh.2023.1249614

Implementation of California COVIDNet – a multi-sector collaboration for statewide SARS-CoV-2 genomic surveillance

Debra A. Wadford[1]*, Nikki Baumrind[1], Elizabeth F. Baylis[1], John M. Bell[1], Ellen L. Bouchard[1], Megan Crumpler[2], Eric M. Foote[1], Sabrina Gilliam[1], Carol A. Glaser[1], Jill K. Hacker[1], Katya Ledin[1], Sharon L. Messenger[1], Christina Morales[1], Emily A. Smith[3], Joel R. Sevinsky[3], Russell B. Corbett-Detig[4†], Joseph DeRisi[5,6‡], Kathleen Jacobson[1‡] and the COVIDNet Consortium

[1]California Department of Public Health, Richmond, CA, United States, [2]Orange County Public Health Laboratory, Santa Ana, CA, United States, [3]Theiagen Genomics, Highlands Ranch, CO, United States, [4]Pathogen Genomics Center, University of California, Santa Cruz, Santa Cruz, CA, United States, [5]University of California, San Francisco, San Francisco, CA, United States, [6]Chan Zuckerberg Biohub, San Francisco, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Ruth Evangeline Timme,
U.S. Food and Drug Administration,
United States

REVIEWED BY
Bernard Joseph Hudson,
New South Wales Health Pathology, Australia
Sarah M. Owens,
Argonne National Laboratory (DOE),
United States

*CORRESPONDENCE
Debra A. Wadford
✉ debra.wadford@cdph.ca.gov

†This author share senior authorship

‡These authors have contributed equally to this work and share last authorship

RECEIVED 29 June 2023
ACCEPTED 27 September 2023
PUBLISHED 23 October 2023

CITATION
Wadford DA, Baumrind N, Baylis EF, Bell JM, Bouchard EL, Crumpler M, Foote EM, Gilliam S, Glaser CA, Hacker JK, Ledin K, Messenger SL, Morales C, Smith EA, Sevinsky JR, Corbett-Detig RB, DeRisi J, Jacobson K and the COVIDNet Consortium (2023) Implementation of California COVIDNet – a multi-sector collaboration for statewide SARS-CoV-2 genomic surveillance.
Front. Public Health 11:1249614.
doi: 10.3389/fpubh.2023.1249614

Introduction: The SARS-CoV-2 pandemic represented a formidable scientific and technological challenge to public health due to its rapid spread and evolution. To meet these challenges and to characterize the virus over time, the State of California established the California SARS-CoV-2 Whole Genome Sequencing (WGS) Initiative, or "California COVIDNet". This initiative constituted an unprecedented multi-sector collaborative effort to achieve large-scale genomic surveillance of SARS-CoV-2 across California to monitor the spread of variants within the state, to detect new and emerging variants, and to characterize outbreaks in congregate, workplace, and other settings.

Methods: California COVIDNet consists of 50 laboratory partners that include public health laboratories, private clinical diagnostic laboratories, and academic sequencing facilities as well as expert advisors, scientists, consultants, and contractors. Data management, sample sourcing and processing, and computational infrastructure were major challenges that had to be resolved in the midst of the pandemic chaos in order to conduct SARS-CoV-2 genomic surveillance. Data management, storage, and analytics needs were addressed with both conventional database applications and newer cloud-based data solutions, which also fulfilled computational requirements.

Results: Representative and randomly selected samples were sourced from state-sponsored community testing sites. Since March of 2021, California COVIDNet partners have contributed more than 450,000 SARS-CoV-2 genomes sequenced from remnant samples from both molecular and antigen tests. Combined with genomes from CDC-contracted WGS labs, there are currently nearly 800,000 genomes from all 61 local health jurisdictions (LHJs) in California in the COVIDNet sequence database. More than 5% of all reported positive tests in the state have been sequenced, with similar rates of sequencing across 5 major geographic regions in the state.

Discussion: Implementation of California COVIDNet revealed challenges and limitations in the public health system. These were overcome by engaging in novel partnerships that established a successful genomic surveillance program which provided valuable data to inform the COVID-19 public health response in California. Significantly, California COVIDNet has provided a foundational data framework and computational infrastructure needed to respond to future public health crises.

KEYWORDS

SARS-CoV-2, genomic surveillance, COVID-19, whole genome sequencing, cloud-based computing, data management

1. Introduction

In early 2020, as SARS-CoV-2 began to spread rapidly around the world, it became clear that an unprecedented, multi-faceted, and coordinated response would be required for this public health crisis. In April 2020, the Governor of California established the California Testing Task Force (CA-TTF)[1] (1) to address the daunting need to provide COVID-19 testing for the state's population of nearly 40 million. Akin to actions taken by the United Kingdom (2), the CA-TTF implemented the California SARS-CoV-2 Whole Genome Sequencing (WGS) Initiative, a genomic surveillance program created to track evolution of the virus over time, monitor variant and lineage transmission throughout the state, and characterize outbreaks and clusters of this virus. Objectives of this program included developing a network of public health laboratories (PHLs) with long-term sequencing capabilities, building genomic epidemiology capability at the state and LHJs for real-time public health action, and establishing and maintaining long-term partnerships among LHJs, academic institutions, and other public, non-profit, and private institutions. This endeavor, named "California COVIDNet," is led by the California Department of Public Health (CDPH) with guidance, support, and input from the Chan Zuckerberg Biohub (CZB) and comprises an exceptional, collaborative network of public, private, and academic laboratories that partnered to scale WGS in response to the COVID-19 pandemic. Herein, we describe the implementation of California COVIDNet, hereafter designated as COVIDNet.

2. Materials and methods

2.1. Implementation of COVIDNet

Implementing COVIDNet required establishing systems and processes, many of which did not exist prior to the pandemic, to manage sample data and sample flow (Figure 1). This included (1) data management systems to anonymize and store SARS-CoV-2 positive sample data to ensure patient privacy, (2) cloud-based storage capacity for WGS data, (3) bioinformatics capabilities for sequence analysis, (4) a network of testing sites and laboratories to source random, representative SARS-CoV-2 positive samples throughout the state, and (5) a network of laboratories to process and sequence samples. Also required were the infrastructure, processes, and procedures to receive sequence data from partner laboratories for centralized and standardized bioinformatics processing, quality control, and analyses for transmission of high-quality lineage results to the state's COVID-19 reporting system and uploading of the data to public repositories. Consultants and contractors with demonstrated expertise in viral evolution, bioinformatics, genomics, and privacy were engaged to advise and guide the execution of many of these processes, including (1) an expert advisory group (see below), (2) a CDPH legal and privacy advisory team, (3) Theiagen Genomics (Highlands Ranch, CO United States), to provide bioinformatics support, including pipelines for data quality control and analysis, lineage reporting and uploading, and cloud-based data storage, and (4) the University of California Santa Cruz (UCSC) Pathogen Genomics Center to develop and implement a system of tools for genomic analyses, including applications for health departments to characterize clusters and outbreaks in their jurisdictions.

2.2. Expert panel advisory group

In June 2020, CDPH convened an advisory panel of nationally and internationally recognized experts comprised of 18 distinguished researchers and subject matter experts in genomic sciences, viral evolution, and mathematical modeling. This expert panel was assembled to support, advise, and guide the initial phases of COVIDNet implementation as well as to engage with LHJs in analyzing and interpreting WGS results.

2.3. Sequencing and data storage capacity

In the first year of the pandemic, most public health-oriented sequencing of California SARS-CoV-2 positive samples was performed by the CZB in association with California local PHLs, as well as by the SEARCH Alliance in San Diego. Although COVIDNet was conceived in April 2020, scaled-up laboratory operations of COVIDNet did not begin until March 2021. This delay was due to factors related to establishing protocols and practices for specimen and data acquisition, flow, and management for a large and populous state. A considerable amount of time and effort was required to recruit and onboard partner laboratories to source or sequence specimens in the midst of the pandemic. CDPH determined and applied best practices to maintain compliance related to data security and privacy, as well as navigate the challenges of logistics and contracting to implement the computational infrastructure required for bioinformatics analytics. CDPH contracted with Theiagen Genomics, which provided a cloud-based solution to house and analyze viral sequence data via Terra.bio on the Google Cloud Platform (GCP).

2.4. Sampling representativeness

One of the goals of genomic surveillance is to ensure sufficient representativeness of the tested population. Successful genomic surveillance of COVID-19 in California required a sequencing strategy representative of the state's geography and diverse

[1] https://testing.covid19.ca.gov/

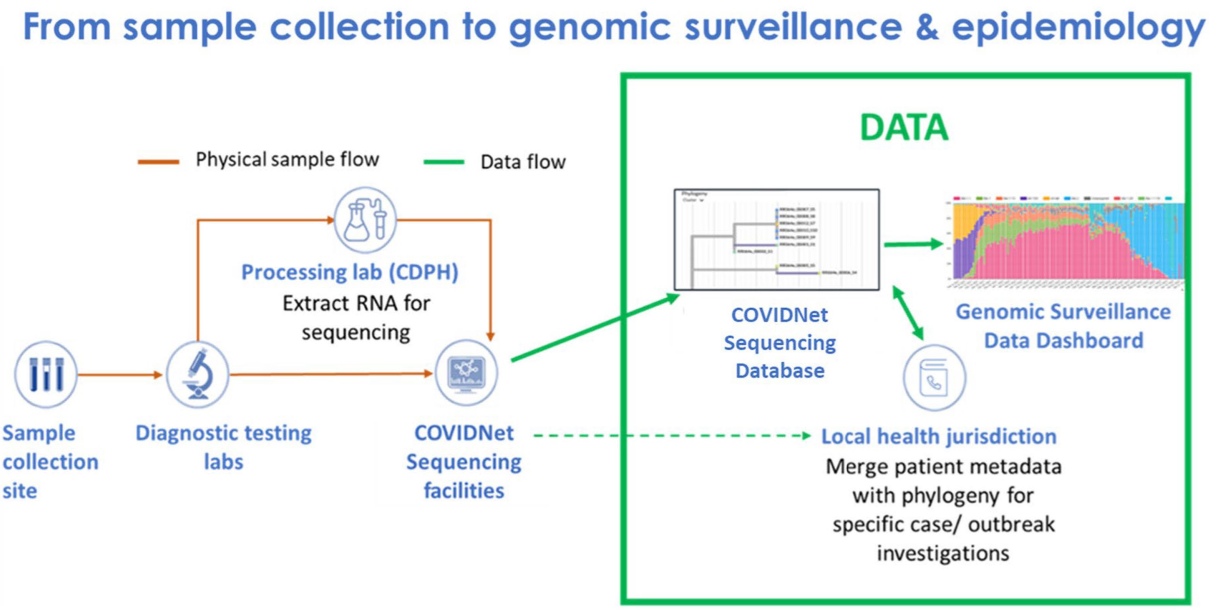

From sample collection to genomic surveillance & epidemiology
Physical sample flow
Data flow
Processing lab (CDPH)
Extract RNA for sequencing
Sample collection site
Diagnostic testing labs
COVIDNet Sequencing facilities
DATA
COVIDNet Sequencing Database
Genomic Surveillance Data Dashboard
Local health jurisdiction
Merge patient metadata with phylogeny for specific case/ outbreak investigations

FIGURE 1
Flow of samples and data from sample collection to genomic sequencing and epidemiology for SARS-CoV-2 genomic surveillance by COVIDNet.

communities that would yield an accurate estimate of SARS-CoV-2 lineages circulating within the state. Achieving the volume and distribution of sequencing that matched true infection rates in each community was not realistic or feasible due to resource and logistical constraints. To coordinate the public health response related to COVID-19 policies (such as stay-at-home orders, tracking hospitalization and ICU capacity, etc.) across the 61 California LHJs, five Health Officers regions were organized as follows: Rural Association of Northern California Health Officers (RANCHO), Association of Bay Area Health Officials (ABAHO), Greater Sacramento Region of Health Officers (GSRHO), San Joaquin Valley Consortium of Health Officers (SJVHO), and Southern California Health Officers (SCHO) (Figure 2) (3). An initial recommendation of the COVIDNet Expert Panel Advisory Group was to target 2 to 5% of the SARS-CoV-2 positive samples in California for sequencing, provided that sampling was random and that less heavily populated counties, such as those in the RANCHO region, were well-represented. To gauge representativeness across the state, we compared sequencing rates from each Health Officers region on a per 100,000-person basis.

2.5. California COVIDNet laboratory network

Success of COVIDNet required establishing a network of diagnostic and sequencing laboratories with varied and critical roles. CDPH developed a Memorandum of Understanding form entitled the "COVIDNet Laboratory Participation Agreement," whereby participating laboratories became official COVIDNet partners to serve as diagnostic, processing (for viral nucleic acid extraction), or sequencing laboratories. Public, private, and academic laboratories became part of this network.

2.6. Scaling sequencing capacity and sample storage

Early in the pandemic, CZB provided free COVID-19 testing and SARS-CoV-2 WGS services, technical consultation, and bioinformatic resources to CDPH, LHJs, and local PHLs. The WGS data generated from CZB served as decisive proof-of-concept that a comprehensive genomic surveillance program for SARS-CoV-2 in California could support and inform public health action and policy. Soon thereafter, the University of California Office of the President (UCOP) and CDPH partnered to quickly establish contracts with eleven UC laboratories to provide WGS capacity services to scale sequencing to at least 5,000 genomes per week. Adding to this capacity were several private laboratories also contracted to provide sequencing services. CDPH established an onsite high-throughput workflow to receive and process thousands of SARS-CoV-2 positive samples every week, extracting and transporting SARS-CoV-2 viral RNA to COVIDNet laboratory partners for sequencing. The standardized extraction protocol, in addition to the standardized analytic pipeline (described below), supported quality control comparisons among the different sequencing methods used by the contracted WGS laboratories.

Aliquots of SARS-CoV-2 samples sent out for sequencing were archived for long-term storage at-80°C. We had to purchase ten additional-80°C freezers to accommodate storage capacity needs due to the large influx of samples during Delta and Omicron surges.

2.7. Centralized repository for sequencing data and designating variants

As previously described (4), CDPH established a centralized sequence repository (COVIDNet sequence database) and analysis structure for SARS-CoV-2 data using cloud storage and computation

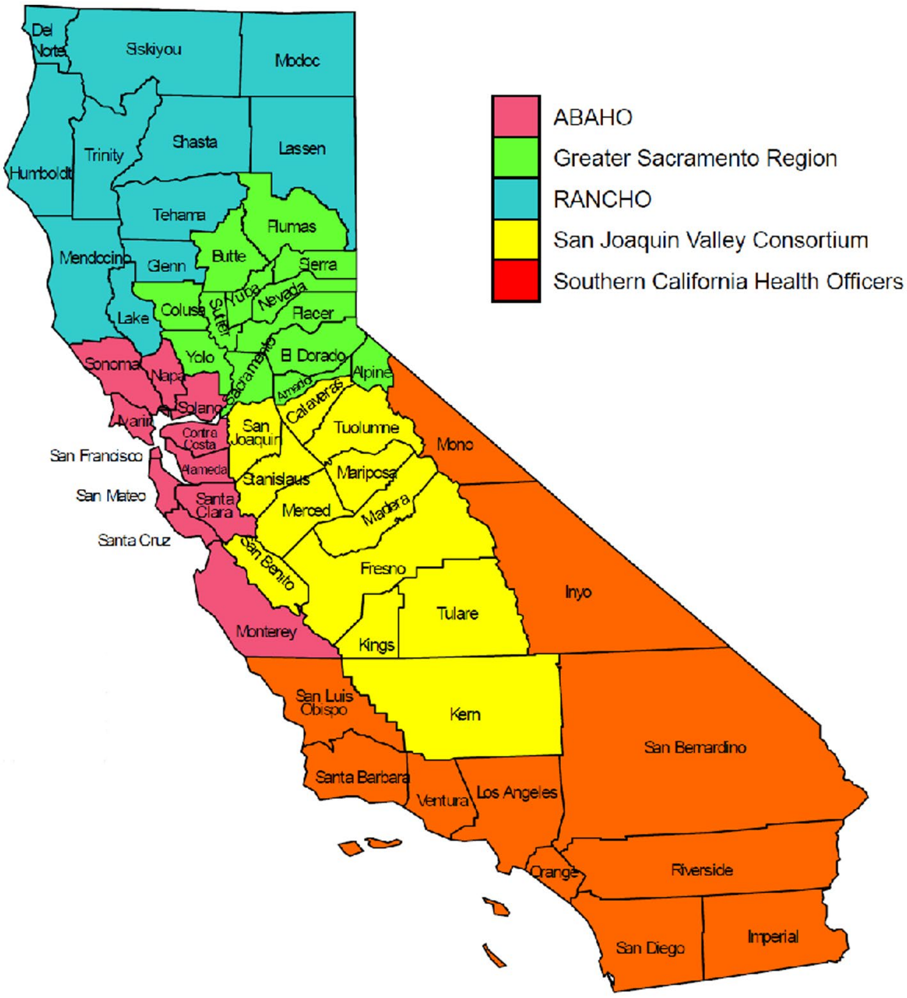


FIGURE 2
The five California Health Officers Regions. Red: Association of Bay Area Health Officials (ABAHO); Green: Greater Sacramento Region of Health Officers (GSRHO); Blue: Rural Association of Northern California Health Officers (RANCHO); Yellow: San Joaquin Valley Consortium of Health Officers (SJVCHO); Orange: Southern California Health Officers (SCHO). https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Order-of-the-State-Public-Health-Officer-Hospital-Health-Care-System-Surge-FAQ.aspx (3).

capabilities (Terra.bio on the GCP) in line with recommendations outlined by Black et al. (5). To ensure CDPH access to SARS-CoV-2 sequence data generated from samples within the state but sequenced by non-COVIDNet partners, the California Code of Regulations (CCR) Title 17 Section 2505 was updated in July 2021 to require that SARS-CoV-2 lineage/variant results including Global Initiative on Sharing All Influenza Data or GISAID (6–8) reference number or raw sequence data, from any California-sourced positive specimen, be reported to CDPH by the sequencing laboratory[2] (9). Sequencing laboratories were provided with an SFTP route to upload FASTQ files and metadata securely to the COVIDNet sequence database and separate metadata repository, respectively. The majority of non-COVIDNet sequence data was submitted to the COVIDNet sequence database by laboratories under contract with the U.S. Centers for Disease Control and Prevention (CDC). Sequence data in the COVIDNet sequence database were processed using a standardized workflow established by Theiagen Genomics, as described by Smith et al. (4). After processing, assembled genomes were uploaded to the public repositories, GISAID or National Center for Biotechnology Information (NCBI) (10).

2.8. Development of analytic tools to support epidemiologic analysis

The UCSC Pathogen Genomics Center was contracted to develop specialized genomic data analytic tools such as on-demand comprehensive phylogenetic resources for CDPH and LHJs, allowing rapid identification and tracking of variants and mutations of interest. UCSC deployed its Ultrafast Sample placement on Existing tRee (UShER) framework (11) to add, in near real-time, every newly sequenced sample from the COVIDNet sequence database to a global phylogenetic tree representing all available genome sequence data

[2] https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/LabReportableDiseases.pdf

from public repositories (currently more than 16 million SARS-CoV-2 genomes). The global phylogenetic tree serves as the source of sequence data to build the California Big Tree – a collection of all California-sourced SARS-CoV-2 sequence data represented in a phylogenetic format[3] (12). UCSC also developed Cluster Tracker, a geo-genomic visualization tool to predict origins of identified genetic clusters[4] (13). This is an exploratory tool that enables users to identify introductions of SARS-CoV-2 into California (state-level, or for authorized users, county-level) and track the geographic clustering of specific SARS-CoV-2 variants or lineages. The tool displays geographic region, sample count within the cluster, clade, lineage, specimen identifiers, and timeframe of clusters based on date of specimen collection. The tool calculates metrics such as a growth score, and best potential origins and indices. Significantly, the user can click on a link to access another UCSC tool called Big Tree Investigator[5] (14), which uses NextStrain (15) to build a phylogenetic tree around a selected cluster and enables linking of California sequences to comprehensive patient-level data reported through the California Reportable Disease Information Exchange (CalREDIE) Electronic Laboratory Reporting (ELR) and non-ELR surveillance systems. These data, displayed together, will enable the authorized CDPH or LHD users to further investigate COVID-19 transmission dynamics within the state and in some cases beyond the state. We expect Big Tree Investigator to go live in December 2023.

2.9. Acquisition, processing, and sequencing of samples

2.9.1. SARS-CoV-2 positive specimens

In general, positive samples detected by molecular methods, including real-time reverse transcription polymerase chain reaction (RT-qPCR), loop-mediated isothermal amplification (LAMP), and transcription-mediated amplification (TMA) assays, were accepted for sequencing. Suitable maximum cycle threshold (Ct) values ranged from 28 to 33. Although Relative Light Unit (RLU) values from transcription mediated amplification (TMA) tests do not correlate directly with viral RNA concentration, TMA specimens were deemed acceptable for WGS if RLU >1,100. Later in the pandemic, as molecular testing rates declined, and antigen testing became more common, reactive swabs from antigen tests were also accepted for sequencing. COVIDNet sequencing prioritized at-risk and vulnerable populations, (e.g., congregate settings such as skilled nursing facilities, prisons, and schools) and known outbreaks, as well as striving to meet equitable representativeness across the state. COVIDNet local PHL partners contributed to these overall goals and many of them prioritized jurisdictional-based investigations of suspected re-infection and vaccine-breakthrough cases and possible importation of new variants from international travelers.

Due to high testing volume and space limitations, many diagnostic laboratories initially discarded SARS-CoV-2 positive specimens before they could be captured for sequencing. To remedy this, in April 2021, CDPH, with the support of LHJs, modified the CCR Title 17 Section 2505 to require diagnostic laboratories to provide COVID-positive remnant specimens to CDPH or a local PHL upon written request[6] (16). Additionally, a centralized COVID-19 testing laboratory, established by the State of California, provided a pipeline of representative COVID-positive specimens collected from more than 7,000 community-based CA-TTF testing sites for WGS. This laboratory aimed to sequence all COVID-positive samples with Ct values less than 33 but ceased operations in May 2022. Additional specimen sources for WGS included those tested by local PHLs that were either sequenced onsite or by a COVIDNet sequencing partner laboratory.

To maintain compliance with California regulations related to personally identifiable information (PII) and protected health information (PHI), samples were de-identified and assigned a 9-or 10-digit Patient Anonymized Unique Identifier (PAUI) to serve as sample identification for sequence data to be processed in the Terra.bio cloud-based platform, uploaded to public repositories such as GISAID and NCBI, and subsequently linked with epidemiologic information in a secure PII/PHI-compliant environment. The PAUI numbers were coded such that the first digit corresponded to particular sample sources or projects in order to distinguish community surveillance samples from samples collected for high priority sequencing or outbreak investigations.

2.9.2. Sample processing and whole genome sequencing

Samples received by CDPH (in either viral transport medium or molecular transport medium) were processed using the KingFisher Flex Purification System (Thermo Fisher Scientific, Waltham, MA United States) nucleic acid extraction platform. Briefly, the lysis step was performed within a class II biological safety cabinet by adding 275 μL of lysis solution containing binding solution and magnetic beads to 200 μL of sample in a 96-well deep well plate. After 10 min of lysis/binding, the plate was loaded onto the KingFisher® instrument along with wash plates, tip comb, and elution plate. Extracted nucleic acid was eluted in 50 μL of elution buffer. Extracts were stored at-70°C and shipped on dry ice to COVIDNet sequencing partners on a weekly basis. Samples that were tested by the LAMP method (Color Health, Burlingame, CA United States) were extracted at the testing facility as follows: total RNA was extracted using the Chemagic® 360 automated system (Perkin-Elmer, Waltham, MA United States). Samples were resuspended in 950 mL lysis buffer (CMG-832). 300uL of lysate was mixed with 300uL of 1x PBS and 10uL of polyA (CMG-842/CMG-843) and 150 uL of magnetic beads (CMG-7000), extracted per manufacturer's protocol, eluted in 80uL of Nuclease-Free Water[7] (17), and shipped on dry ice to CDPH as nucleic acid extracts. The State's COVID-19 testing laboratory performed nucleic acid extractions using the Chemagic® extraction platform as described above[8] (18). COVIDNet partner PHLs that performed their own sequencing followed nucleic acid extraction protocols compatible with their sequencing protocols.

[3] https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

[4] https://clustertracker.gi.ucsc.edu/

[5] https://pathogengenomics.ucsc.edu/tools

[6] https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/LabReportableDiseases.pdf

[7] https://www.fda.gov/media/138249/download

[8] https://www.fda.gov/media/147547/download

TABLE 1 SARS-CoV-2 sequencing protocols employed by California COVIDNet WGS laboratory partners.

Sequencing protocol	Number of COVIDNet laboratories
ARTIC or modified ARTIC[a]	27
Clear Dx[b]	11
Swift[c]	3
Varskip[d]	2

[a]ARTIC v.3, v.4, and v.4.1 (19, 20), ARTIC v.3 with 275 bp tailed amplicons (21).
[b]Clear Dx (22).
[c]SWIFT protocol (23).
[d]Varskip (24).

Because of the varying capabilities and instrumentation available among the COVIDNet sequencing laboratory partners, library preparation and sequencing protocols used were dependent upon the particular sequencing method employed by the laboratory (Table 1). Sequencing library preparation methods included standard ARTIC v.3, v.4, and v.4.1 (19, 20), ARTIC v.3 with 275 bp tailed amplicons[9] (21), the SWIFT protocol (23), Midnight protocol (25), and Varskip (24). Sequencing technologies included Illumina (San Diego, CA United States) MiSeq®, Illumina NextSeq®, Illumina NovaSeq®, AVITI™ (Element Biosciences, San Diego, CA United States), and Clear Dx (San Carlos, CA United States) (22). Illumina sequencing included single-end as well as paired-end protocols.

2.10. Data management: processing, analysis, and storage

Since it was not feasible to standardize the library preparation and sequencing methods by the various COVIDNet contributors, homogeneity of analysis was achieved by having all sequence data centralized and analyzed using a standardized workflow. Uniform workflow, bioinformatics analytics, and training resources were established (4) utilizing Terra.bio[10] (26), as the centralized location to house COVIDNet sequence data.

Raw sequence data reads, in FASTQ file format, were made available through various methods on Terra.bio. County PHLs shared FASTQ files from other cloud-based platforms such as Illumina's BaseSpace (San Diego, CA) and the Clear Labs Portal (San Carlos, CA). Read data hosted on Illumina's BaseSpace platform were made available on Terra.bio through the BaseSpace_Fetch workflow[11] (27), while reads stored on the Clear Labs Portal were made available on Terra.bio directly through the portal's user interface. Academic COVIDNet partners were provisioned with GCP buckets to serve as persistent storage. They uploaded FASTQ files directly to these buckets, and the data stored in these GCP buckets were made accessible on Terra.bio through an automated cron job which ran once daily. Alternatively, FASTQ files were manually uploaded from a local machine to Terra.bio.

Once the FASTQ files were available on Terra.bio, genome assembly and characterization were performed using the TheiaCov Workflows for Genomic Characterization. This open-source workflow series is available on the Public Health Viral Genomics Github repository[12] (28) and includes workflows for analysis of Illumina paired-end, Illumina single-end, Oxford Nanopore (Oxfordshire, England), Element Biosciences AVITI (San Diego, CA), and Clear Labs SARS-CoV-2 data. In addition to assembling the genome, these workflows also provided quality control metrics, Pango lineages (29), and Nextclade clades (15). The StaPH-B docker image for Pangolin[13] (30, 31) was used within the TheiaCov workflow, wherein the UShER (11) mode of Pangolin was used by default for lineage assignment. Whenever the docker image was updated following a pangolin-data[14] (32) release, the Pangolin_Update workflow[15] (33) on Terra.bio was run to assign updated lineages to California sequences with collection dates in the past two months.

SARS-CoV-2 sequence data from CDPH were submitted to GISAID and NCBI if the genome assembly covered at least 83% of the Wuhan-1 reference genome (MN908947) (34), as determined by the TheiaCov workflows. The Mercury workflow series on Terra.bio, also hosted within the Public Health Viral Genomics Github repository (28) were used to reformat the FASTA files and metadata according to the submission guidelines for GISAID and NCBI. Genome assemblies from CDPH were uploaded to GISAID and GenBank (35) and raw reads were uploaded to the Sequence Read Archive (SRA) (36) with the exception of data generated on the Element Biosciences AVITI instrument, as data from that instrument cannot be accepted at this time. Raw reads uploaded to CDPH SARS-CoV-2 BioProject (PRJNA750736) were depleted of host reads using the SRA human read scrubber[16] (37). All California local PHLs that used Terra.bio also uploaded to GISAID and NCBI using their own quality control thresholds for submission. It is important to note that samples sequenced by local PHLs were distinct from the samples sequenced by COVIDNet sequencing laboratories and therefore duplicate submissions to public repositories were highly unlikely and not considered a concern. The Terra.bio platform also allowed local PHLs to customize and optimize workflows for their own use, such as to build automated import and export pipelines to decrease reliance on manual processes.

Data on Terra.bio are exported to external Google buckets for downstream visualization, alerting, and reporting using the Terra_2_BQ workflow[17] (38). Data were then ingested into Big Query projects using either a cron job running a shell script, or a Google workflow. An SQL query was used to combine data from all COVIDNet partners into a single data source. Looker, a GCP for automated alerts and reports, or Looker Studio, a visualization platform, was used to monitor the changes in SARS-CoV-2 lineages over time.

[9] https://www.protocols.io/view/ucsf-cat-covid-19-tailed-275bp-v3-artic-protocol-v-kxygxpnpzl8j/v1

[10] https://terra.bio

[11] https://github.com/theiagen/terra_utilities/blob/main/workflows/wf_basespace_fetch.wdl

[12] https://github.com/theiagen/public_health_viral_genomics

[13] https://hub.docker.com/r/staphb/pangolin/

[14] https://github.com/cov-lineages/pangolin-data

[15] https://github.com/theiagen/public_health_viral_genomics/blob/main/workflows/wf_pangolin_update.wdl

[16] https://github.com/ncbi/sra-human-scrubber

[17] https://github.com/theiagen/terra_utilities/blob/main/workflows/wf_terra2bq.wdl

TABLE 2 SARS-CoV-2 genomes stored in the California COVIDNet sequence database (March 2020 through March 2023).

Laboratory Category	Total genomes sequenced	Total genomes uploaded* and accepted#	Percent of genomes uploaded* and accepted #
COVIDNet contract	356,875	263,706	74%
Public Health Laboratory (State and local)	93,155	81,131	87%
Totals	450,030	344,837	77%

*Genomes that achieved ≥ 83% reference coverage were uploaded to GISAID or NCBI sequence repositories. #A genome was considered accepted by GISAID if it was released upon curation without requiring further confirmation or modification of the sequence data.

2.11. Quality of sequence data

Successful sequencing was determined by "percent reference coverage," i.e., the proportion of the genome successfully sequenced to a minimum depth of 20X. The minimum percent reference coverage required for uploading to a public sequence repository was 83%, approximately 25,000 bases of a SARS-CoV-2 genome. Other quality metrics assessed, but not necessarily used to reject sequences, included: (1) the quantiles of sequencing depth across the results of a sequencing run; (2) the proportion of putative human DNA (as found via Kraken2 (39) analysis) vs. the percent reference coverage; and, (3) sample or within-run contamination, as evidenced either by high proportions of apparent minor allele frequency single nucleotide polymorphisms (SNPs) across samples in a sequencing run or by improbable phylogenetic placement, such as detection of a variant from samples collected prior to the emergence of said variant. More recently (as of April 2023), samples with a disproportionately high number of variants (>17.5%) with allele frequencies between 60 and 90% were rejected due to the likelihood that such samples were either contaminated or coinfected. We assumed that contamination was more likely in these cases than coinfection, and that coinfections, if present, would be rare and unlikely to affect the understanding of surveillance significantly. The lower threshold of 60% was selected because it is also the minimum threshold for calling a variant allele for assembly and thus may have affected lineage determination and phylogenetic placement.

3. Results

One primary goal of COVIDNet was to sequence 2–5% of SARS-CoV-2 positive samples in California, in a representative and equitable manner. Through collaborative agreements and contracts with a variety of partners, we successfully engaged 50 laboratories including public health, academic, and private laboratories to achieve large-scale WGS of SARS-CoV-2 throughout California. As of March 2023, CDPH has received more than 660,000 samples from various submitters throughout the state and processed and extracted more than 217,000 samples (data not shown) that met WGS criteria (e.g., Ct value <33 or from reactive antigen tests). As the pandemic progressed and the COVIDNet workflow became routine, we expanded SARS-CoV-2 genomic surveillance to include (1) testing sites along the international border between California and Mexico, (2) at three international airports, (3) at community organizations serving priority populations, and (4) at all schools participating in the CA-TTF testing program. In mid-June 2021, we partnered with a large integrated health system that serves over 4.5 million members in Northern and Central California (40) to characterize SARS-CoV-2 variants in both inpatient and outpatient populations, further expanding the COVIDNet surveillance network. In the latter part of 2022, as antigen testing began to predominate over molecular testing for SARS-CoV-2, select COVIDNet partners transitioned WGS operations to accept swabs from reactive antigen tests to maintain an adequate level of surveillance as much as possible.

Through COVIDNet, WGS capacity and bioinformatics capability increased within the network of 29 California PHLs at the state and local levels. By the end of 2021, a total of 15 (52%) of 29 PHLs were conducting SARS-CoV-2 genomic surveillance via COVIDNet. To date, WGS capacity for SARS-CoV-2 has been established at 19 of the 29 (66%) PHLs in the state. Six California PHLs have hired bioinformaticians for SARS-CoV-2 sequence analysis. In mid-2022, CDPH formally established a new Genomic Epidemiology Section to analyze, manage, and apply SARS-CoV-2 genomic surveillance data for situational awareness, and to inform infectious disease modeling and forecasting[18] (41), public health action, and policy.

3.1. Sequencing volume

As shown in Table 2, between March 2020 through March 2023, a total of 450,030 genomes were deposited by COVIDNet partner laboratories into the California COVIDNet sequence database in Terra.bio, with 344,837 genomes (77%) meeting the 83% reference coverage threshold to upload to GISAID or NCBI. The percent of genomes uploaded was higher overall for PHLs (87%) than for COVIDNet contract laboratories (74%) (Table 2). In some cases, this was likely due to differences in specimen handling. Samples processed at the COVIDNet contract laboratories typically underwent several freeze-thaw cycles prior to extraction, had to be transported to CDPH for extraction, or were of poor quality. The percentage of sequences assigned to a lineage increased over time and has remained above 80% since July 2021 (Figure 3). From March 2020 through March 2023, CDC-contracted laboratories contributed more than 335,000 sequences to the California COVIDNet sequence database (data not shown). Most of the California-sourced samples (~75%) sequenced by CDC-contracted laboratories were from the Southern California Health Officer (SCHO) region (data not shown), the most populous region of the state (Table 3). Samples sequenced by CDC-contracted laboratories made up 66% of samples from the SCHO region in the

[18] https://calcat.covid19.ca.gov/cacovidmodels/

COVIDNet sequence database, compared to less than 30% from the four other regions (Figure 4). In regions other than the SCHO, more than 50% of the samples were sequenced by COVIDNet laboratory partners (Figure 4).

Between March 2020 and December 2020, an average of 361 samples were sequenced per month, with an average of 0.7% of positive tests sequenced (data not shown). From January 2021 through June 2021, a time period that encompasses the start of COVIDNet sequencing in March 2021, an average of 20,875 samples were sequenced per month, with an average of 7.9% (~11-fold increase from 2020) of positives sequenced. The large-scale genomic surveillance efforts across California by COVIDNet and CDC-contracted laboratories has resulted in nearly 800,000 SARS-CoV-2 genomes in the COVIDNet sequence database. Between March 2020 and March 2023, these data have allowed us to monitor the emergence and spread of different variants statewide including the emergence of Epsilon in the fall and winter of 2020, followed by co-circulation of Alpha and P.1 (Gamma) in the spring of 2021 (Figure 5). Later we observed the transition to Delta in June 2021 followed by the abrupt introduction and predominance of Omicron BA.1 in December 2021 with subsequent diversification of Omicron sublineages, and dominance of BA.5 from June 2022 into January 2023. We saw the rise of the XBB and other recombinants in late winter of 2022 with XBB.1.5 predominating at the end of March 2023 (Figure 5).

The percentage of positive samples sequenced peaked at over 21% in July 2021 (Figure 6). Sequencing volume peaked between July 2021 and February 2022, averaging 52,941 samples per month, which correlated with a spike in the number of reported positive tests in the state. The percent of positives sequenced averaged 7.1% but dropped to a low of 1.6% in January 2022 during the Omicron surge (Figure 6). From March 2022 through January 2023, sequencing volume declined to an average of 16,881 sequences per month, and an average of 6% of positive tests sequenced.

3.2. Sequencing representativeness

All 61 California health jurisdictions were represented within the COVIDNet sequence database. Sequencing representativeness was similar across the five Health Officers regions of California. The number of samples sequenced per 100,000 people ranged from 1,650 to 2,086 from March 2020 through March 2023 (Table 3). This value varied over time, but in total remained similar across Health Officers regions at specific time points, except for a spike in August 2021 from the RANCHO region (Figure 7). The proportion of sequences per respective Health Officers region corresponded approximately well with each region's population percentage (Table 3). Given that nearly two-thirds of the sequences generated by CDC-contracted laboratories were from the SCHO Region, sample representativeness across other regions in the state was assured by supplementing with COVIDNet sequencing of samples from CA-TTF community-based sites and other sources (Figure 4).

3.3. Quality of sequence data

We routinely monitored the quality of sequence data generated by the COVIDNet sequencing laboratories which varied across the network (data not shown). We met with the sequencing laboratories monthly to review overall sequencing quality and success of individual sequencing runs; occasionally we requested re-sequencing when a high proportion of failures or evidence of significant contamination was detected. Although the measure of human DNA vs. the percent reference coverage was not used as a criterion for rejection, low percent reference coverage (i.e., failed sequencing) tended to strongly correlate with a high proportion of human DNA in a sample. While the overall sequencing quality was acceptable across COVIDNet partners (as represented by proportion of sequences assigned a lineage designation) (Figure 3), some laboratories experienced sporadic issues

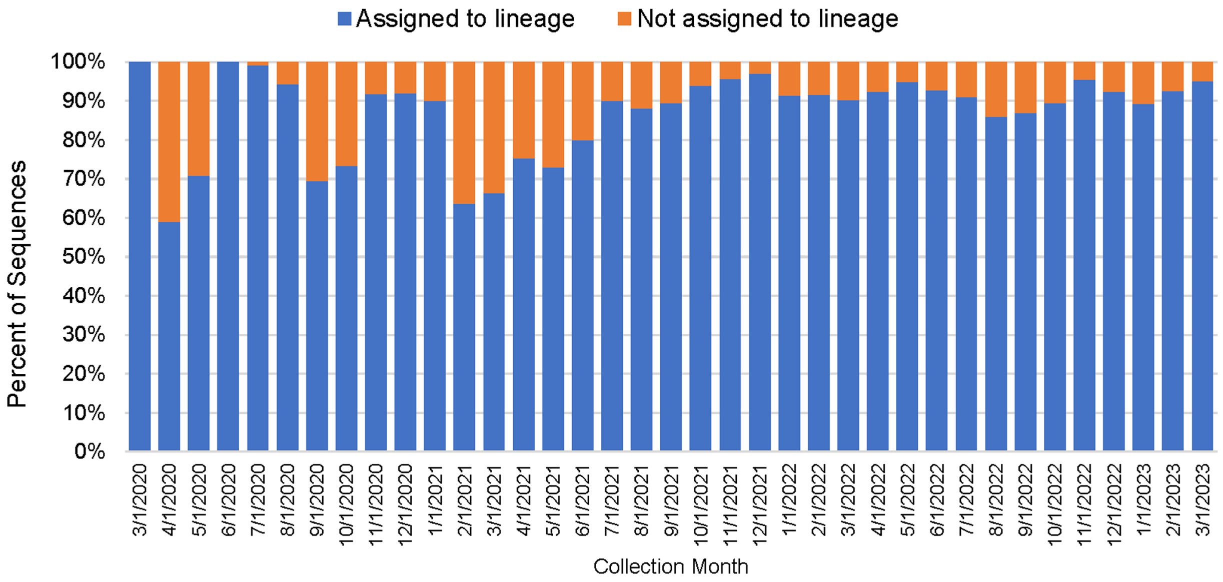


FIGURE 3
Percentage of SARS-CoV-2 sequences in the California COVIDNet sequence database assigned to a lineage or not assigned (March 2020 to March 2023).

TABLE 3 SARS-CoV-2 sequencing volume by California Health Officers Region and population (March 2020 – March 2023).

California (CA) Health Officers region	Population (2021)	Percent of CA population	Number of sequences	Percent of sequences	Number of sequences (per 100,000 population)
ABAHO	8,451,422	21%	163,948	24%	1,940
GSRHO	2,964,755	8%	51,489	7%	1,737
RANCHO	701,548	2%	14,631	2%	2,086
SJVCHO	4,470,528	11%	86,492	13%	1,935
SCHO	22,867,100	58%	377,387	54%	1,650
California Total	39,455,353	100%	693,947	100%	1,783
ABAHO: Association of Bay Area Health Officials					
GSRHO: Greater Sacramento Region of Health Officers					
RANCHO: Rural Association of Northern California Health Officers					
SJVHO: San Joaquin Valley Consortium of Health Officers					
SCHO: Southern California Health Officers					

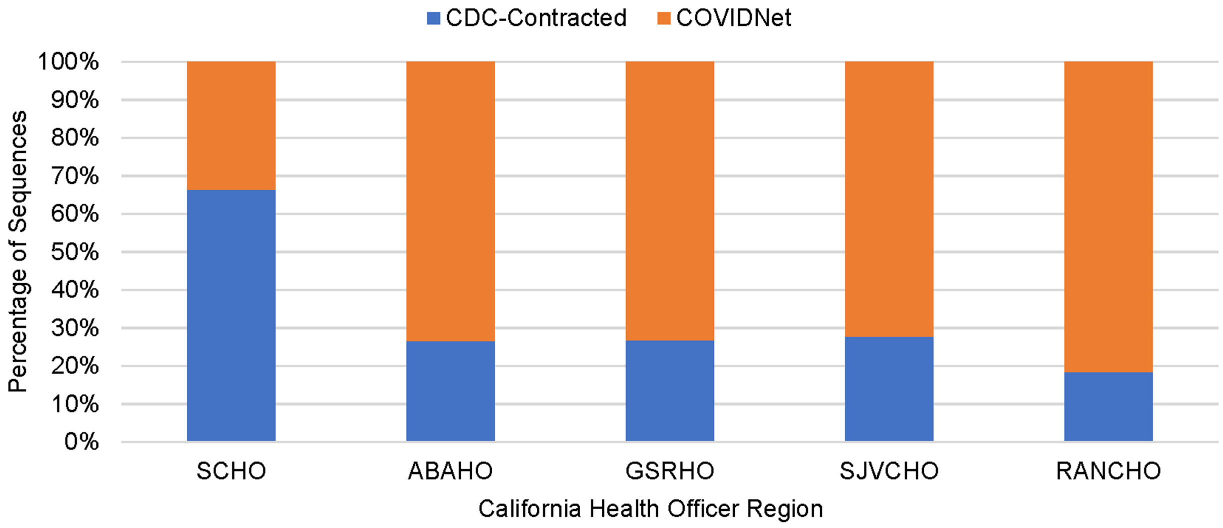


FIGURE 4
Percentage of sequenced samples for each California Health Officers Region by laboratory category (March 2020 to March 2023).

requiring re-sequencing and root-cause analysis to prevent recurrent quality excursions (data not shown). Problems included general quality issues such as unacceptably low numbers of samples passing quality metrics on a given sequencing run, as well as occasional detection of contamination from a sample or within a sequencing run, as evidenced either by high proportions of minor allele frequency SNPs across samples within a sequencing run, or for example, by the improbable detection of putative Omicron sequences from samples collected prior to the known date of Omicron emergence. In some instances, we observed declines in sequence quality over time as the virus evolved which indicated the need to update sequencing protocols or primers, particularly with the emergence of the Omicron variant in November 2021. In general, the goal of data quality monitoring was to keep the number of errors in sequences released to public repositories low, while not slowing throughput needed to maintain relevant and current surveillance.

4. Discussion

4.1. Public health impact of COVIDNet

The large-scale approach to genomic surveillance implemented by California COVIDNet partners and others has enabled CDPH and LHDs to monitor the evolution of SARS-CoV-2 over time including transitions between viral variants of interest and variants of concern (VOC). COVIDNet efforts have provided valuable data supporting epidemiologic investigations and policy-making decisions in California, at the state and local levels. COVIDNet data have revealed local trends of viral transmission, helped to characterize and better understand outbreaks of SARS-CoV-2 within skilled nursing facilities, schools, and other settings (42–50) and have provided situational awareness of circulating variants with the potential to impact efficacy of vaccines and therapeutics. Of particular note was the use of WGS

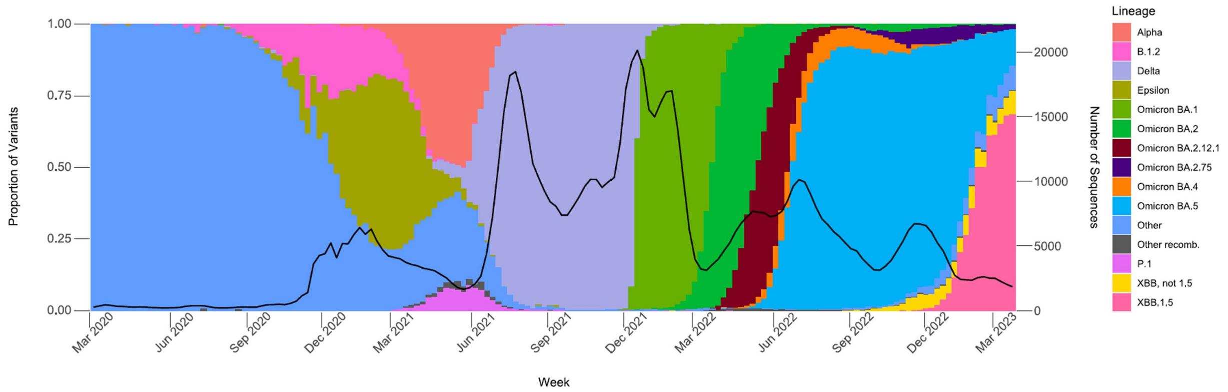


FIGURE 5
Proportions of major SARS-CoV-2 variants by collection date from the California COVIDNet sequence database, March 2020 through March 2023. Data graphed represent weekly proportions with the black line representing the 4-week rolling average number of samples sequenced. Colors correspond to indicated SARS-CoV-2 variant lineage.

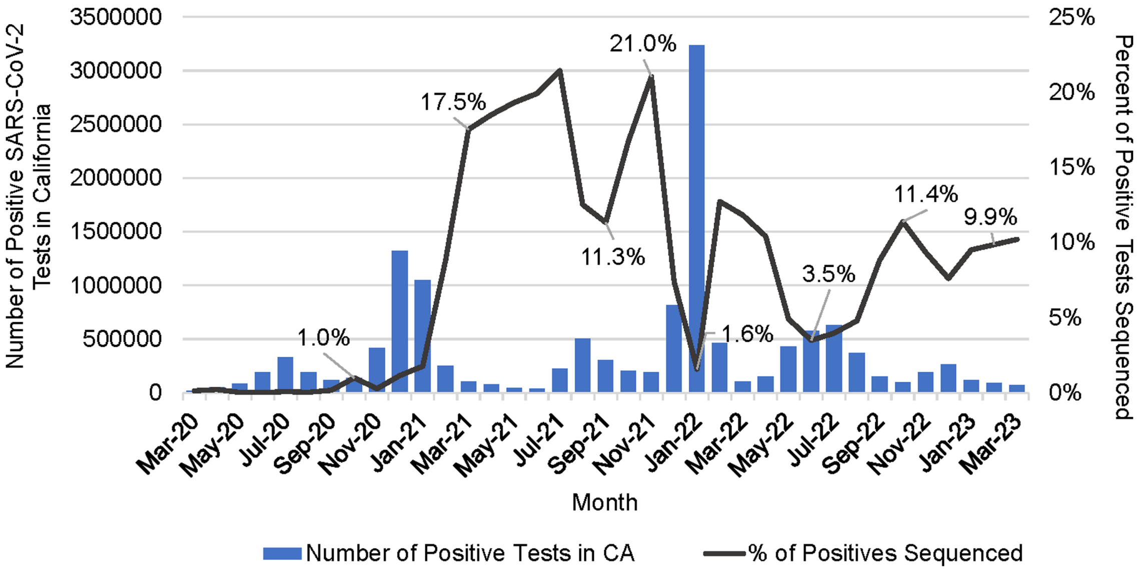


FIGURE 6
Number of positive SARS-CoV-2 tests in California (CA) per month and the percent of positive tests that were sequenced per month (March 2020 to March 2023).

data to assess the levels of circulating SARS-CoV-2 variants with known resistance to monoclonal antibodies that effectively ruled out such treatment. Furthermore, COVIDNet data have enabled identification and characterization of cases and variants associated with vaccine breakthrough infections, the first cases of new VOCs in the state, re-infections (40) (45) (49), and long-term infections demonstrating intra-host evolution (unpublished). Regardless of publication status, sequence data meeting quality criteria have been uploaded in a timely manner to data repositories for public access. COVIDNet efforts have provided data enabling California to establish its own COVID forecasting model available online to the public[19] (41).

The cloud-based data infrastructure developed to store the large volume of COVIDNet WGS data and to provide a framework for analysis has created capability, not only for SARS-CoV-2 but also other pathogens, such as monkeypox virus (MPXV), enteric bacteria, and select pathogens associated with hospital infections. The workflows for SARS-CoV-2 sequence analysis have been leveraged for wastewater surveillance (WWS) applications and will be utilized in the continued expansion of WWS at CDPH. Using automated tools (e.g., Google Looker) to query the COVIDNet sequence database at defined time periods, we set up email alerts to notify relevant public health officials about the detection or emergence of concerning variants, mutations of interest, and proportions of variants detected weekly (4).

In April 2022, as the COVID-19 pandemic continued, the emergence and subsequent global spread of MPXV occurred (51–53).

[19] https://calcat.covid19.ca.gov/cacovidmodels/

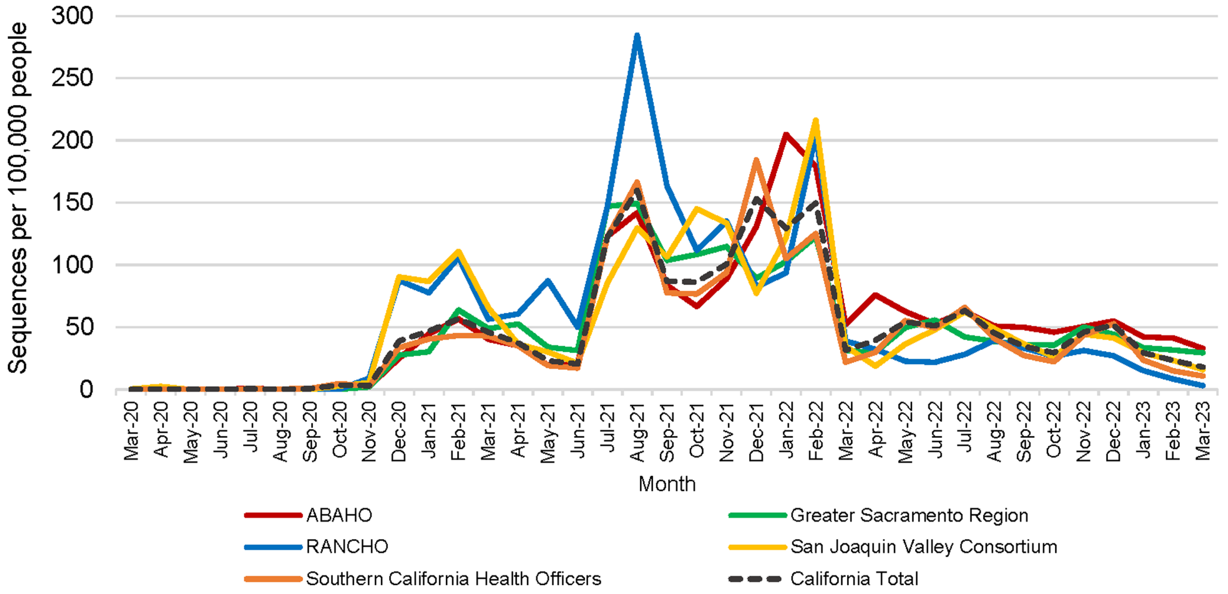


FIGURE 7
Number of SARS-CoV-2 sequences per 100,000 people by month for each California Health Officers region (March 2020 to March 2023).

Although WGS protocols and data analysis for this emerging pathogen were not in place at that time, we were able to modify the SARS-CoV-2 WGS workflow and adapt the COVIDNet data analysis infrastructure for timely MPXV sequence analysis. MPXV WGS results were shared among California LHDs and PHLs to examine transmission patterns and evolution of MPXV as it spread in different regions of the state.

4.2. Accomplishments

COVIDNet successfully achieved many of its original goals and objectives. We achieved large-scale genomic surveillance of SARS-CoV-2 across California, which allowed us to monitor the emergence and spread of variants statewide including the lineage diversification of Omicron variants and the rise of SARS-CoV-2 recombinants (Figure 5). Nineteen of 29 California PHLs have established WGS capability for SARS-CoV-2 and are now thusly prepared for future public health crises and pandemics. Due to the efforts of both COVIDNet partners and non-COVIDNet laboratories, there are nearly 800,000 genomes in the COVIDNet sequence database, and as cases of COVID-19 continue to occur, this number will continue to increase.

With COVIDNet, we established significant collaborations with academic and private partners that strengthened statewide capacity to respond to COVID and future infectious disease threats. We will endeavor to maintain these important partnerships to benefit and ensure preparedness for public health. The success of COVIDNet demonstrates the power of productive collaborations among California's public, private, and academic institutions in responding to an unprecedented international public health emergency. The response to COVID-19 in California laid the foundation for COVIDNet, as a system, to be adapted for other pathogens of public health importance and future public health emergencies.

4.3. Challenges

The accomplishments of implementing COVIDNet did not occur without challenges, some resolved, some still ongoing, with many inherent to outdated public health infrastructure and data systems. Challenges included navigating hierarchies of data management needs from samples, sequences, metadata, and results. Because COVIDNet receives disparate data from multiple sources and projects, we were forced to develop multiple databases to manage incoming data and samples amid viral pandemic surges. This approach allowed us to quickly scale during the pandemic, but it created burdensome processes for managing and analyzing data from the variety of sources. It also caused inconsistencies and redundancies in data, such as duplicate sequences with different internal identifiers, inability to easily manage multiple samples from single individuals, and incomplete metadata.

Patient-level metadata is frequently classified PII and PHI that cannot be housed in the same location as its matching viral genomic data. The COVIDNet sequence database is specifically a PII/PHI-free cloud-based platform, consistent with this tenet. It is a significant challenge to join sequence data with PII/PHI data due to incompatibilities between the two data systems and associated privacy regulations. Substantial investments were made in collaboration with UCSC to develop a tool, Big Tree Investigator, to automate integration of sequence data from the COVIDNet sequence database with patient-level data located in a separate secure PHI-compliant environment. Big Tree Investigator advances beyond earlier genomic epidemiology tools to facilitate linking of these databases to visualize sequence data with associated PHI-metadata mapped on a phylogenetic tree. Such visualizations will provide context from other sequences and metadata to understand clusters and outbreaks to, possibly, contain/control further transmission within a defined community or region (i.e., conduct genomic epidemiology). Developing this tool has been impeded by complications relating to PII/PHI concerns and limited assets, but we anticipate that Big Tree Investigator will go live in December 2023.

4.4. Limitations

The COVIDNet project had several limitations, many of which still exist. First and foremost was the static and risk-averse nature of public health infrastructure and systems at the state and local levels that were insufficient to support advances in sequencing technology and attendant data requirements, particularly during the chaos of the pandemic. Prior to the pandemic, funding for public health had declined and was usually insufficient to implement modern systems or needed updates. The pandemic resulted in increased funding for public health response, but in many cases the infrastructure was a hindrance in putting the funds to use quickly or optimally.

A major limitation was achieving timely generation of sequence data. Contracts with COVIDNet partners established a 2-week turnaround time for sequence results but delays of 2 weeks or more to process and ship samples out to COVIDNet partners were typical, impeding timely results. Although CDPH can provide rapid SARS-CoV-2 WGS (49), this was not scalable and proved useful only for certain situations, such as high-risk outbreaks or severe cases.

Limitations in sampling strategies to meet goals of representativeness and equity were manageable because of widespread availability and accessibility of COVID-19 testing from established community-based testing sites and other sources. The closure of these testing sites, as well as the transition from molecular-based to antigen-based testing made it difficult to maintain the goal of geographic and equitable representation and we expect this challenge to continue. As sources of samples for sequencing decrease, the risk of surveillance bias will also increase. Given the success of SARS-CoV-2 wastewater surveillance (54–56), we expect wastewater surveillance to continue to contribute a significant portion of data to inform SARS-CoV-2 genomic surveillance going forward and that this will help to mitigate surveillance bias.

A further limitation is that, although we established a PAUI system to distinguish between surveillance samples and other high priority/outbreak samples, the COVIDNet database contains sequences that do not have PAUI numbers and thus it is not always clear whether these samples are surveillance-based or from targeted investigations and therefore not randomly selected. Thus, outbreak specimens and specimens from high-risk settings may be over-represented as surveillance data. Likewise, the vast majority of specimens that were sequenced did not include attendant detailed medical or travel history, and thus this genomic surveillance program was primarily laboratory-based.

4.5. Recommendations and conclusions

As we move to the next phase of COVID-19, it is important to ensure funding to support ongoing genomic surveillance for SARS-CoV-2 and other pathogens of public health significance globally, nationally, and at state and local levels. Secure and sustained funding is critical to ensure the capacity to identify and analyze emerging pathogens of concern. Resources are also required to address gaps in public health infrastructure, in particular, systems related to data management, storage, and analysis. These systems are in significant need of modernization and functional interoperability to optimize data transmissions.

To build on experience gained with COVIDNet and prepare for future crises, it is imperative that we create, in advance, uniform data requirements with the ability to integrate incoming data from many sources, and transition to database and query platforms that can handle very large datasets.

Partnerships between public health departments and clinical laboratories should be established to help enhance collaborations and prepare us for the next large outbreak or pandemic. By enhancing these partnerships, we can help to identify and submit specimens for genomic surveillance as part of routine public health surveillance systems and also provide critical genomic information for the medical community that might help guide care and treatment. Protocols should be developed between these institutions to create a sample referral system in the event of outbreaks of concern or public health emergencies so that specimens are not discarded before they can be captured for sequencing or other characterization.

The value of COVIDNet will continue with ongoing genomic surveillance of SARS-CoV-2, as well as other pathogens of public health significance. Importantly, the COVIDNet infrastructure has already demonstrated, with the emergence of MPXV, its utility to be leveraged for the next public health crisis. The flexibility of this large-scale, collaborative system provides the necessary methods, logistics, and workflows that require only minor modifications to enable effective genomic surveillance and epidemiology at local, regional, and state levels.

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 at: https://www.ncbi.nlm.nih.gov/, Bioproject number PRJNA750736.

Group members of the COVIDNet Consortium

Summer Adams, Phacharee Arunleung, Matthew Bacinskas, Nikki Baumrind, Elizabeth F. Baylis, Cynthia Bernas, John M. Bell, Ricardo Berumen, Ellen L. Bouchard, Brandon Brown, Teal Bullick, Lyndsey Chaille, Alice Chen, Giorgio Cosentino, Yocelyn Cruz, Nick D'Angelo, Mojgan Deldari, Alex Espinosa, Ambar Espinoza, Eric M. Foote, Gautham, Shiffen Getabecha, Sabrina Gilliam, Carol A. Glaser, Madeleine Glenn, Bianca Gonzaga, Ydelita Gonzales, Melanie Greengard, Hugo Guevara, Jill K. Hacker, Kim Hansard, April Hatada, Monica Haw, Thalia Huynh, Kathleen Jacobson, Chantha Kath, Paul B. Kimsey, Katya Ledin, Deidra Lemoine, Ruth Lopez, Sharon L. Messenger, Blanca Molinar, Christina Morales, Samantha Munoz, Robert Nakamura, Nichole Osugi, Tasha Padilla, Chao-Yang Pan, Mayuri V. Panditrao, Chris Preas, Will Probert, Alexa Quintana, Maria Uribe-Fuentes, Mayra Ramirez, Clarence Reyes, Estela Saguar, Maria Salas, Ioana Seritan, Brandon Stavig, Hilary Tamnanchit, Serena Ting, Debra A. Wadford, Cindy Wong, Chelsea Wright, and Shigeo Yagi, California Department of Public Health (CDPH); Venice Servellita, Alicia Sotomayor-Gonzalez, and Charles Y. Chiu, Chiu laboratory at University of California, San Francisco; Isabel Bjork, Joshua Kapp, Anouk van den Bout, and Ellen Kephart, Colligan Clinical Diagnostic Lab, University of California, Santa Cruz; Mawadda Alnaeeli, Hau-Ling Poon, Scott Topper, Color Health; Marzieh Shafii, Sara Sowko, Stephanie Trammell, and Erik Wolfsohn, Contra Costa County PHL; Patrick Ayscue, Amy Kistler, Emily Crawford, Joseph DeRisi, and Cristina Tato, Chan Zuckerberg Biohub; Valeria Arboledaz, Eleazar Eskin, and Laila M. Sathe, Department of Computational Medicine, University of California, Los Angeles; Jacek Skarbinski, Kaiser Permanente Northern California, Division of Research; Abigail Duque, Jeffrey Schapiro, and Ivy Yeung, Kaiser Permanente Northern California, The Permanente Medical Group; Rama Ghatti and Zahra Shajani-Yi, LabCorp; Jacob M. Garrigues, Nicole Green, and Peera Hemarajata, Los Angeles County PHL; Carlos Anaya and Donna Ferguson, Monterey County PHL; Beatrix Kapuszinsky, Favian Ramirez, and Felipe Sta Agueda, Napa-Solano-Yolo-Marin-Mendocino County PHL; Megan Crumpler and Julia Wolfe, Orange County PHL; Russell B. Corbett-Detig, David Haussler, Marc Perry, and Jakob McBroome, Pathogen Genomics Center, University of California, Santa Cruz; Nhi Duong, Deborah Forester, and Anthony Gonzalez, Sacramento County PHL; Maria J. Victorio, Anna Liza M. Manlutac, Jeremy Corrigan, and Nicholas S. Rhoades, San Diego County PHL; Lina Castro and Godfred Masinde, San Francisco County PHL; Harmeet Kaur, Monica Paniagua-Alexander, San Joaquin County PHL; Katrina G. Erwin, Glen Miller, and Frances N. Sidhu, San Luis Obispo County PHL; Morris Jones, Sangita Kothari, and Christopher Ngo, San Mateo County PHL; Brandon Bonin, Daniel Castillo, and Rensen Khoshabian, Santa Clara County PHL; Kristian Andersen, Mark Zeller, SEARCH Alliance; Lisa Critchett, Carlos Gonzalez, Iryna V. Goraichuk, and Rachel Rees, Sonoma County PHL; Frank Ambrosio, Curtis J. Kapsak, Kevin G. Libuit, Michelle R. Scribner, Joel R. Sevinsky, Emily A. Smith, and Sage M. Wright, Theiagen Genomics; Vanessa B. Cadiz, Denise Lopez, and Matthew Rosman, Tulare County PHL; Bryan Bach, Stacia Wyman, UC Berkeley Innovative Genomics Institute SARS-CoV-2 Sequencing Team; Charlotte Acharya, Ryan Davis, and Richard Michelmore, UC Davis Genome Center; Melanie Oakes and Suzanne Sandmeyer, UC Irvine Campus Covid Testing Laboratory & Genomics High Throughput Facility; Kathy Borkovich, Clay H. Clark, Holly Clark, and Brandon Le, UC Riverside Institute for Integrative Genome Biology - Genomics Core Facility; Peter De Hoff, Kristen Jepsen, Rob Knight, and Louise C. Laurent, UC San Diego EXCITE Laboratory; Zack Aralis and Carolina Arias, UC Santa Barbara Arias Lab; Varuzhan Balasanyan, Mark Duhon, and Xinmin Li, UCLA Technology Center for Genomics & Bioinformatics; Eric Chow, Nicole Leung, Delsy Martinez, and Tyler T. Miyasaki, UCSF Center for Advanced Technology; Ashlee Clow, Jared Hoffman, and Thomas Rush, Ventura County PHL.

Author contributions

DW, JD, SM, KJ, JB, and RC-D conceived of this project, and guided it to fruition. DW, JB, NB, SG, ES, SM, JH, EF, ELB, CG, KL, and KJ wrote and revised the manuscript. All members of the COVIDNet Consortium contributed to data acquisition, analysis, results interpretation, analysis, results interpretation, or supervision of the work. All authors contributed to the article and approved the submitted version.

Funding

CDPH/COVIDNet genomic surveillance work was funded by the Centers for Disease Control and Prevention, Epidemiology and Laboratory Capacity for Infectious Diseases, Cooperative Agreement Number 5 NU50CK000539. Sequencing work by the Chiu laboratory at University of California, San Francisco was funded by US CDC award 75D301-21-C-10991 (CYC).

Acknowledgments

We appreciate the confidence and support of Dr. Charity Dean and Paul Markovich, inaugural Co-Chairs of the California Testing Task Force, which enabled us to make California COVIDNet a reality. We are grateful to Lifang Chiang, Theresa Maldonado, and Brian Russ from the University of California Office of the President for their assistance in facilitating collaborations with University of California laboratories. We thank Gary Chang, Sarah Folker, Lisa Gorog, Courtney Hansen, Bridget Jones, Elaine Owen, and Jill Tellioglu from the CDPH Office of Legal Services. We sincerely appreciate the guidance from the COVIDNet Expert Advisory Panel members Kristian Andersen, The Scripps Research Institute; Patrick Ayscue, Chan Zuckerberg Biohub; Josh Batson, previously with Chan Zuckerberg Biohub; John M. Bell, CDPH; Sidney M. Bell, Chan Zuckerberg Initiative; Shannon N. Bennett, California Academy of Sciences; Charles Chiu, University of California, San Francisco; Russ B. Corbett-Detig, University of California, Santa Cruz; David Dynerman, previously with Chan Zuckerberg Biohub; David Haussler, University of California, Santa Cruz; A. Marm Kilpatrick, University of California, Santa Cruz; Amy Kistler, Chan Zuckerberg Biohub; Ha Youn Lee, University of Southern California; Joseph Lewnard, University of California, Berkeley; Duncan MacCannell, US Centers for Disease Control and Prevention; Ayesha Mahmud, University of California, Berkeley; Richard Michelmore, University of California, Davis; Erin A. Mordecai, Stanford University; Julia A. Palacios, Stanford University; Joel R. Sevinsky, Theiagen Genomics; Marc A. Suchard, University of California, Los Angeles; Paul D. Thomas, University of Southern California; George G. Vega Yon, University of Southern California; Katharine S. Walter, Stanford University (current institution University of Utah); Joel Wertheim, University of California, San Diego. We thank the following CDPH colleagues for their assistance and guidance: Tomas Aragon, Allison Bailey, Briyana Baker, Tricia Blocher, Deva Borthwick, Tim Bow, Cathy Catholic, Holli Dowless Kozar, Jesse Elder, Susan Fanelli, Lidia Gomes, Carl V. Hanson, Kathleen Harriman, Zoe Hawk-Pratty, Seema Jain, Andrew Knapp, Dominick Morales, Mary Kate Morris, Erica S. Pan, Peter Patiris, Eric Piano, Melissa Relles, Rahil Ryder, Angela Salas, Pat Stoll, Anthony Tran, Wilfred Velasco, John Walker, and James P. Watt. We are grateful for the following colleagues and entities: Vici Varghese, Alameda County PHL; Greg Costo and Regina Gallick Bertolucci, Butte County PHL; Allison Black, previously with Chan Zuckerberg Initiative; Jeffrey Bulawit, Fresno Co PHL; Pepper Stockton, Humboldt County PHL and Shasta County PHL; Holly Maag, Imperial County PHL; Amparo Sembrano, Long Beach PHL; Si Noon Hong, Madera Co PHL; Carol Barrera, Ana Gonzales, and Cheryl Starling, Heluna Health; Neena Bhakta and Errin Rider, Riverside County PHL; Linda Ward and Jessica Valdez, San Bernardino County PHL; Katayoon Dehesh, UC Riverside Institute for Integrative Genome Biology; Ashka Dave, CA-TTF; Adam Dondro, Secretary Mark Ghaly, and Marko Mijic, California Health and Human Services Agency; Kimberley Lucas, California Correctional Health Care Services; Ephrem Chin and Madhuri Hegde, CDPH Valencia Branch Laboratory/PerkinElmer; Catelyn Andersen, Alison King, Ezra Kurzban, Kelly Nguyen, Sarah Perkins, Karthik Ramesh, Kieran Saucedo, Madison Schwab, and Alana Weiss of the SEARCH Alliance; Avellino Laboratories, Avrok Laboratories, Curative, Fulgent Genetics, Invitae, Kaiser Permanente Northern California Regional Laboratory, LabCorp, LetsGetChecked Laboratories, and Quest Diagnostics.

Conflict of interest

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

Publisher's note

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

Author disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the California Department of Public Health or the California Health and Human Services Agency.

References

1. California Department of Public Health. California Coronavirus COVID-19 Testing Task Force: CA Web Publishing Service (2023). Available at: https://testing.covid19.ca.gov/

2. Aanensen DM, Khalil A, Adams A, Afifi S, Alam MT, Alderton A, et al. Coronavirus disease 2019 (COVID-19) genomics UK consortium. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet. *Microbe*. (2020) 1:e99–e100. doi: 10.1016/S2666-5247(20)30054-9

3. California Department of Public Health. Public Health Order Questions & Answers: Hospital & Health Care System Surge (2021). Available at: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Order-of-the-State-Public-Health-Officer-Hospital-Health-Care-System-Surge-FAQ.aspx

4. Smith EA, Libuit KG, Kapsak CJ, Scribner MR, Wright SM, Bell J, et al. Pathogen genomics in public health laboratories: successes, challenges, and lessons learned from California's SARS-CoV-2 whole-genome sequencing initiative, California COVIDNet. *Microbial Genomics*. (2023) 9:001027. doi: 10.1099/mgen.0.001027

5. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat Med*. (2020) 26:832–41. doi: 10.1038/s41591-020-0935-z

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

7. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. (2017) 22:30494–5. doi: 10.2807/1560-7917.ES.2017.22.13.30494

8. GC KS, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, et al. GISAID's role in pandemic response. *China CDC Wkly*. (2021) 3:1049–51. doi: 10.46234/ccdcw2021.255

9. California Department of Public Health. Title 17, California Code of Regulations (CCR), Section 2505 Subsection (q), page 6 (2022). Available at: https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/LabReportableDiseases.pdf

10. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. (2022) 50:D20–6. doi: 10.1093/nar/gkab1112

11. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet*. (2021) 53:809–16. doi: 10.1038/s41588-021-00862-7

12. University of California SCGBg. UShER: Ultrafast Sample placement on Existing tRee (2022). Available at: https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

13. University of California SCGI. (2021). Available at: https://clustertracker.gi.ucsc.edu/.

14. University of California SCGI. UCSC Pathogen Genomics Tools (2021). Available at: https://pathogengenomics.ucsc.edu/tools

15. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407

16. California Department of Public Health. Title 17, California Code of Regulations (CCR), Section 2505 Subsection (p), page 5. (2022). Available at: https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/LabReportableDiseases.pdf

17. Hinton DM. *U.S. Food and Drug Administration. Emergency use authorization (EUA) summary for the color SARS-CoV-2 Rt-lamp diagnostic assay*. (2021). Available at: https://www.fda.gov/media/138249/download

18. U.S. Food and Drug Administration. Emergency Use Authorization (EUA) Summary PerkinElmer SARS-CoV-2 RT-qPCR Reagent Kit (PerkinElmer Genomics). (2021). Available at: https://www.fda.gov/media/147547/download

19. Quick J, Loman NhCoV-2019/nCoV-2019 Version 3 Amplicon Set: ARTIC Consortium; (2020). Available at: https://artic.network/resources/ncov/ncov-amplicon-v3.pdf

20. Loman Nartic-ncov2019 primer schemes: ARTIC Consortium; (2021). Available at: https://github.com/artic-network/primer-schemes

21. University of California SF. *UCSF CAT COVID-19 Tailed 275bp v3 ARTIC protocol* v1 V.1. (2021) Available at: https://www.protocols.io/view/ucsf-cat-covid-19-tailed-275bp-v3-artic-protocol-v-kxygxpnpzl8j/v1

22. Ramaiah A, Khubbar M, Scott S, Bauer A, Lentz J, Akinyemi K, et al. Implementation and evaluation of the clear dx platform for sequencing SARS-CoV-2 genomes in a public health laboratory. *Microbiology Spectrum*. (2023) 11:e04957-22.

23. Addetia A, Lin MJ, Peddu V, Roychoudhury P, Jerome KR, Greninger AL. Sensitive recovery of complete SARS-CoV-2 genomes from clinical samples by use of Swift Biosciences' SARS-CoV-2 multiplex amplicon sequencing panel. *J Clin Microbiol*. (2020) 59. doi: 10.1128/JCM.02226-20

24. New England Biolabs Inc. VarSkip multiplex PCR designs for SARS-CoV-2 sequencing. *GitHub*; Available at: https://github.com/nebiolabs/VarSkip. (2022)

25. Freed NE, Vlkova M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore rapid barcoding. *Biol Methods Protoc*. (2020) 5:bpaa014. doi: 10.1093/biomethods/bpaa014

26. Theiagen Genomics. (2023) Terra. Available at: https://terra.bio

27. Theiagen Genomics Git Hub. (2021). Available at: https://github.com/theiagen/terra_utilities/blob/main/workflows/wf_basespace_fetch.wdl

28. Theiagen Genomics. Public Health Viral Genomics. (2021). Available at: https://github.com/theiagen/public_health_viral_genomics GitHu

29. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. (2020) 5:1403–7. doi: 10.1038/s41564-020-0770-5

30. Oakeson K, Wang S, Florek K, Fink L, Hanigan C, MacCannell D. *State public health bioinformatics community. Staphb/pangolin - software package for assigning SARS-CoV-2 genome sequences to global lineages Docker hub*. (2020). Available at: https://hub.docker.com/r/staphb/pangolin/

31. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus evolution*. (2021) 7:veab064. doi: 10.1093/ve/veab064

32. Cov-Lineages. cov-lineages/pangolin-data. (2023). Available at: https://github.com/cov-lineages/pangolin-data.GitHub

33. Theiagen Genomics. Pangolin Update. (2023). Available at: https://github.com/theiagen/public_health_viral_genomics.GitHu

34. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. (2020) 579:265–9. doi: 10.1038/s41586-020-2008-3

35. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, EW OJS. GenBank. *Nucleic Acids Res*. (2013) 41:D36–42. doi: 10.1093/nar/gks1195

36. Leinonen R, Sugawara H, Shumway MInternational Nucleotide Sequence Database C. The sequence read archive. *Nucleic Acids Res*. (2011) 39:19–21. doi: 10.1093/nar/gkq1019

37. NCBI SRA human read scrubber. (2023). Available at: https://github.com/ncbi/sra-human-scrubber.GitHub

38. Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, et al. (2023). Terra_2_BQ workflow Available at: https://github.com/theiagen/terra_utilities/blob/main/workflows/wf_terra2bq.wdl.GitHub

39. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*. (2019) 20:1–13. doi: 10.1186/s13059-019-1891-0

40. Skarbinski J, Nugent JR, Wood MS, Liu L, Bullick T, Schapiro JM, et al. SARS-CoV-2 Delta variant genomic variation associated with breakthrough infection in northern California: a retrospective cohort study. *J Infect Dis*. (2023):jiad164). doi: 10.1093/infdis/jiad164

41. California Department of Public Health. California Communicable diseases Assessment Tool (2023). Available online at: https://calcat.covid19.ca.gov/cacovidmodels/

42. Karmarkar EN, Blanco I, Amornkul PN, DuBois A, Deng X, Moonan PK, et al. Timely intervention and control of a novel coronavirus (COVID-19) outbreak at a large skilled nursing facility—San Francisco, California, 2020. *Infection Control & Hospital Epidemiology*. (2021) 42:1173–80. doi: 10.1017/ice.2020.1375

43. Villarino E, Deng X, Kemper CA, Jorden MA, Bonin B, Rudman SL, et al. Introduction, transmission dynamics, and fate of early severe acute respiratory syndrome coronavirus 2 lineages in Santa Clara County. *California J infectious dis*. (2021) 224:207–17. doi: 10.1093/infdis/jiab199

44. Deng XGW, Federman S, du Plessis L, Pybus OG, Faria NR, Wang C, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into northern California. *Science*. (2020) 369:582–7. doi: 10.1126/science.abb9263

45. Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cells*. (2021) 184:3426–37. doi: 10.1016/j.cell.2021.04.025

46. Lam-Hine T, McCurdy SA, Santora L, Duncan L, Corbett-Detig R, Kapusinszky B, et al. Outbreak associated with SARS-CoV-2 B. 1.617. 2 (delta) variant in an elementary school—Marin County, California, may–June 2021. *Morbidity and mortality weekly report*. (2021) 70:1214. doi: 10.15585/mmwr.mm7035e2

47. Stoddard G, Black A, Ayscue P, Lu D, Kamm J, Bhatt K, et al. Using genomic epidemiology of SARS-CoV-2 to support contact tracing and public health surveillance in rural Humboldt County, California. *BMC Public Health*. (2022) 22:456. doi: 10.1186/s12889-022-12790-0

48. MacCannell T, Batson J, Bonin B, Astha KC, Quenelle R, Strong B, et al. Genomic epidemiology and transmission dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in congregate healthcare facilities in Santa Clara County. *California Clinical Infectious Diseases*. (2022) 74:829–35. doi: 10.1093/cid/ciab553

49. Wadford DA, Page L, Mosack K, Bauer A, Khubbar M, Balakrishnan N, et al. *Poster: Application of rapid whole genome sequencing to identify a SARS-CoV-2 omicron variant outbreak among highly vaccinated wedding attendees*. OH, Cleveland: Association of Public Health Laboratories 2022 Annual Conference. (2022).

50. Lam-Hine T, McCurdy SA, Santora L, Duncan L, Corbett-Detig R, Kapusinszky B, et al. *Outbreak associated with SARS-CoV-2 B. 1.617. 2 (delta) variant in an elementary school—Marin County, California, may–June 2021. Morbidity and mortality weekly report*. (2021), 70:1214.

51. Thornhill JP, Barkati S, Walmsley S, Rockstroh J, Antinori A, Harrison LB, et al. Monkeypox virus infection in humans across 16 countries—April–June 2022. *N Engl J Med*. (2022) 387:679–91. doi: 10.1056/NEJMoa2207323

52. Minhaj FS, Ogale YP, Whitehill F, Schultz J, Foote M, Davidson W, et al. Monkeypox outbreak—nine states, may 2022. *Morb Mortal Wkly Rep*. (2022) 71:764. doi: 10.15585/mmwr.mm7123e1

53. Isidro J, Borges V, Pinto M, Sobral D, Santos JD, Nunes A, et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. *Nat Med*. (2022) 28:1569–72. doi: 10.1038/s41591-022-01907-y

54. Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, Lou YC, et al. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *MBio*. (2021) 12:10–128. doi: 10.1128/mBio.02703-20

55. Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature*. (2022) 609:101–8. doi: 10.1038/s41586-022-05049-6

56. Li L, Uppal T, Hartley PD, Gorzalski A, Pandori M, Picker MA, et al. Detecting SARS-CoV-2 variants in wastewater and their correlation with circulating variants in the communities. *Sci Rep*. (2022) 12:16141. doi: 10.1038/s41598-022-20219-2

frontiers | Frontiers in Cellular and Infection Microbiology

TYPE Original Research
PUBLISHED 31 October 2023
DOI 10.3389/fcimb.2023.1251456

Whole genome analysis of *Rhizopus* species causing rhino-cerebral mucormycosis during the COVID-19 pandemic

Joy Sarojini Michael[1]*, Manigandan Venkatesan[1], Marilyn Mary Ninan[1], Dhanalakshmi Solaimalai[1], Lydia Jennifer Sumanth[1], Lalee Varghese[2], Regi Kurien[2], Rinku Polachirakkal Varghese[3] and George Priya Doss C[3]

[1]Department of Clinical Microbiology, Christian Medical College, Vellore, Vellore, Tamil Nadu, India, [2]Department of Otorhinolaryngology, Christian Medical College, Vellore, Vellore, India, [3]Department of Integrative Biology, School of Biosciences and Technology, Vellore Institute of Technology (VIT) University, Vellore, Tamil Nadu, India

OPEN ACCESS

EDITED BY
Shangxin Yang,
University of California, Los Angeles, United States

REVIEWED BY
Jayapradha R.,
SASTRA University, India
Travis K. Price,
Quest Diagnostics, United States
J. R. Caldera,
UCLA Health System, United States

*CORRESPONDENCE
Joy Sarojini Michael
✉ joymichael@cmcvellore.ac.in

RECEIVED 01 July 2023
ACCEPTED 11 September 2023
PUBLISHED 31 October 2023

CITATION
Michael JS, Venkatesan M, Ninan MM, Solaimalai D, Sumanth LJ, Varghese L, Kurien R, Varghese RP and C GPD (2023) Whole genome analysis of *Rhizopus* species causing rhino-cerebral mucormycosis during the COVID-19 pandemic.
Front. Cell. Infect. Microbiol. 13:1251456.
doi: 10.3389/fcimb.2023.1251456

Introduction: Mucormycosis is an acute invasive fungal disease (IFD) seen mainly in immunocompromised hosts and in patients with uncontrolled diabetes. The incidence of mucormycosis increased exponentially in India during the SARS-CoV-2 (henceforth COVID-19) pandemic. Since there was a lack of data on molecular epidemiology of Mucorales causing IFD during and after the COVID-19 pandemic, whole genome analysis of the Rhizopus spp. isolated during this period was studied along with the detection of mutations that are associated with antifungal drug resistance.

Materials and methods: A total of 50 isolates of *Rhizopus* spp. were included in this prospective study, which included 28 from patients with active COVID-19 disease, 9 from patients during the recovery phase, and 13 isolates from COVID-19-negative patients. Whole genome sequencing (WGS) was performed for the isolates, and the *de novo* assembly was done with the Spades assembler. Species identification was done by extracting the ITS gene sequence from each isolate followed by searching Nucleotide BLAST. The phylogenetic trees were made with extracted ITS gene sequences and 12 eukaryotic core marker gene sequences, respectively, to assess the genetic distance between our isolates. Mutations associated with intrinsic drug resistance to fluconazole and voriconazole were analyzed.

Results: All 50 patients presented to the hospital with acute fungal rhinosinusitis. These patients had a mean HbA1c of 11.2%, and a serum ferritin of 546.8 ng/mL. Twenty-five patients had received steroids. By WGS analysis, 62% of the *Rhizopus* species were identified as *R. delemar*. Bayesian analysis of population structure (BAPS) clustering categorized these isolates into five different groups, of which 28 belong to group 3, 9 to group 5, and 8 to group 1. Mutational analysis revealed that in the *CYP51*A gene, 50% of our isolates had frameshift mutations along with 7 synonymous mutations and 46% had only synonymous mutations, whereas in the *CYP51B* gene, 68% had only synonymous mutations and 26% did not have any mutations.

Conclusion: WGS analysis of Mucorales identified during and after the COVID-19 pandemic gives insight into the molecular epidemiology of these isolates in our community and establishes newer mechanisms for intrinsic azole resistance.

KEYWORDS

whole genome sequencing, molecular epidemiology, Mucorales, COVID - 19, azole resistance detection

Introduction

Mucorales are common environmental molds that cause mucormycosis. This is an opportunistic fungal infection that is angio-invasive and therefore has high morbidity and mortality. Even though mucormycosis is found worldwide, causative agents are more common in India. Among the order Mucorales, *Rhizopus arrhizus*/*R. oryzae* is the most common species isolated in the laboratory, followed by *Rhizopus microspores, Litchthemia, Cunningamella*, and *Saksena*. Patients with diabetes mellitus, hematological malignancy and chemotherapy, and hematopoietic stem cell transplant, and solid organ transplant recipients on immunosuppressive therapy with iron overload are at risk of developing mucormycosis. The most common clinical presentation is invasive fungal sinusitis or rhino-orbital–cerebral mucormycosis (ROCM), followed by pulmonary, gastrointestinal, cutaneous, and renal mucormycosis.

The delta wave of the pandemic swept through India from May 2021. There was an increase in the incidence of mucormycosis in patients with SARS-CoV-2 (henceforth COVID-19) during this wave around the world, particularly from India. Epidemiological reviews reveal an acute increase in the incidence of ROCM related to COVID-19 infection. Phylogeny of Mucorales isolated during the COVID-19 pandemic has not been studied in India, where many cases were reported, including from our center (Cherian et al., 2022).

Before the COVID-19 pandemic, the death rate for mucormycosis was 50%; however, during the delta wave, fatalities amounted to 85% (Aranjani et al., 2021). Owing to the rise in mucormycosis cases during this wave of the COVID-19 pandemic and its link with fatalities in COVID-19 patients, further studies on mucormycosis are needed particularly to investigate the relationship of Mucorales with COVID-19 patients (Al-Tawfiq et al., 2021).

So far, genotyping of Mucorales has been performed by using the internally transcribed spacer (ITS) region and D1/D2 regions of the 28S rRNA subunit (Nagao et al., 2005), or multilocus sequencing typing of conserved loci (Cendejas-Bueno et al., 2012). These methods do not reflect genome-scale phylogenetic differences adequately or correctly capture strain and species-level diversity. Whole genome sequencing (WGS) has been used in the recent studies to investigate mucormycosis outbreaks. Though it has inherent challenges, WGS analysis will help to understand the biology and pathogenesis of the organism and disease.

Azoles inhibit ergosterol synthesis by interacting with the 14-∝ sterol demethylases, encoded in molds by CYP51 genes. Azole resistance in filamentous fungi are due to overexpression of CYP51A and/or point mutations in the CYP51A gene and overexpression of efflux pumps. Macedo et al. in 2018 describe that in *Rhizopus oryzae*, CYP51 genes are uniquely responsible for intrinsic resistance to short-tailed triazoles such as voriconazole and fluconazole (Macedo et al., 2018).

Therefore, in this study, we performed WGS on 50 isolates of *Rhizopus* spp. isolated during the delta wave of the COVID-19 pandemic from COVID-positive, -recovered, and -negative patients. We wanted to ascertain the phylogenetic relationship among the isolates in these three groups and to study whether evolutionary clusters and the presence of mutations in the CYP51 genes played a role in the severity of the disease in COVID-19 patients.

Materials and methods

Ethics

This study was approved by the Christian Medical College, Vellore Institutional Review board and Ethics committee (IRB no. 14007).

Study population and sample collection

This was a prospective study done at Christian Medical College Vellore, a large tertiary care teaching hospital that saw many patients with COVID-19-associated mucormycosis. Consecutive clinical isolates of *Rhizopus arrhizus* were cultured from patients with ROCM during the delta wave of the COVID-19 pandemic between March 2021 and December 2021 to collect isolates from post-COVID-19 and COVID-19-negative patients. All isolates were retrieved from patients presenting with a rhino-orbital cerebral sinusitis and were from the sinus tissue; some had extensions into the brain and some into the bone and orbit. The sinus tissue samples obtained from these patients were minced with sterile scissors in a sterile petri dish. The presence of sparsely septate broad, irregular hyphae branching at obtuse angles on microscopic calcofluor white microscopic preparation was identified, and it was cultured on Sabourauds Dextrose Agar with and without antibiotics as per standard laboratory procedure. Characteristic features on culture and microscopy identified the cultures. COVID-19 testing at our center was carried out by the Altona Realstar SARS-CoV-2 RTPCR kit and the Cepheid Xpert Xpress SARS-CoV-2 assay. Of the 50 isolates collected during this period, 28 belonged to patients with active COVID-19 disease (within 3 weeks of RT-PCR positivity), 9 were from patients in the recovery phase (after 3 weeks of RT-PCR positivity), and 13 isolates were from COVID-19-negative patients (negative RT-PCR test).

DNA extraction

Mucorale isolates were stored at room temperature and subcultured onto Sabouraud Dextrose Agar before processing for DNA extraction. Once grown on Sabouraud Dextrose Agar, they underwent Genomic DNA extraction using the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) per the manufacturer's instructions. Good-quantity and -quality DNA was selected, and WGS was further carried out. The isolated DNA was quantified using a QubitTM 3 Fluorometer (Thermo Fisher Scientific) and a minimum of 0.3 ng/μL DNA concentration was required to perform WGS. DNA quality was verified by running agarose gel electrophoresis to detect nucleic acid degradation. Extracted DNA was stored at −20°C until further use.

Whole genome sequencing

KAPA HyperPrep Kit (Roche) was used to prepare Illumina sequencing libraries according to the manufacturer's instructions. After preparing the DNA sample libraries, they were purified with Ampure XP Reagent (Beckman Coulter), quantified with 5300 Fragment Analyzer (Agilent), and uniquely barcoded multiple samples libraries were normalized together to be sequenced equally and simultaneously in a single run. Then, the libraries were sequenced with a 2×150-bp paired-end reads chemistry on the Illumina NovaSeq platform as per the manufacturer's instructions, resulting in an average of 100× coverage of the whole genome per isolate for all samples.

Genome assembly

Sequence reads were trimmed to remove poor-quality bases using Trimmomatic (v0.39) followed by *de novo* assembly with Spades (v3.14.1) with the following k-mer lengths: 27, 33, 55, and 75.

Species tree generation based on ITS gene

Mucorales, in comparison with other genetic targets like 18S and D1/D2 of the 28S gene ITS region, shows higher species-specific variability and may further discriminate species in *Rhizopus* species (Nagao et al., 2005). Abe et al. also describe better clustering of isolates using ITS region sequencing (Abe et al., 2006). Based on this, ITS gene sequences were extracted from our isolates and reference isolates genome with the BLASTN 2.12.0+ tool and the combined sequence of ITS1-5.8S-ITS2 genes were searched as a query in the Nucleotide BLAST database (https://blast.ncbi.nlm.nih.gov/) for species identification. Genetic clustering analysis for our isolates along with reference isolates was done with RhierBAPS 1.1.4 (Tonkin-Hill et al., 2018). ITS gene sequence multiple alignment was done using Geneious software (https://www.geneious.com/) with Clustal Omega v1.2.3 (Sievers and Higgins, 2018) for our isolates along with the reference isolates, and species tree was generated with RAxML 8.2.11 with the following parameters: GTRGAMMA nucleotide model, Rapid Bootstrapping and a search for best-scoring maximum likelihood tree, 100 bootstrap replicates with parsimony random seed 100, followed by visualization and annotation using iTOL (https://itol.embl.de/) (Figure 1).

Based on work from Macedo et al. (2018), two types of CYP51 genes were identified in the *R. oryzae* genome, and they were classified as CYP51A and CYP51B (Macedo et al., 2018). *R. oryzae* ATCC 11886 CYP51A and CYP51B gene sequences were downloaded from the NCBI database and blasted with our isolate genome with the help of the BLASTN 2.12.0+ tool; conversion of the resulting nucleotide sequence to protein sequence followed by multiple alignment with reference protein sequence was done using Geneious software (https://www.geneious.com/) with Clustal Omega v1.2.3 and studied for mutations.

Phylogenetic tree construction using eukaryotic reference marker genes

Our clinical isolate-assembled genomes were analyzed with the Benchmarking Universal Single-Copy Orthologs (BUSCO v5.4.4) tool to look for orthologous groups specific to the fungi_odb10 lineage using Augustus (v3.3) as described by Simao et al. and Stanke et al (Simão et al., 2015). (Stanke et al., 2006) and the following parameter: "*Rhizopus oryzae*" species was selected for genome assessment mode and protein-coding genes were predicted. Predicted protein sequences were extracted from the AUGUSTUS output. The PhyloSift reference marker genes for eukaryotes as described by Darling et al. (2014) were downloaded and concatenated into one combined file for query and then searched against our isolates that predicted protein sequences (Nguyen et al., 2020). From the 33 reference marker genes found to be conserved among all eukaryotic organisms, only 12 of them were present with complete sequence across 30 of our clinical isolate genomes; thus, they were utilized to compare the phylogenetic relatedness between those isolates (Figure 2). Those are 14_3_3, Actin_noOuts, Atub_noOuts, Btub_noOuts, enolase, gamma_noOuts, hsp70, hsp70cyt, hsp70er, Rps23a_noOuts, TFIIH, and U5. Nucleotide sequences of these marker genes from our isolates were aligned using Geneious software (https://www.geneious.com/) with Clustal Omega v1.2.3 as described by Sievers and Higgins et al (Sievers and Higgins, 2018). to make a combined multiple alignment file

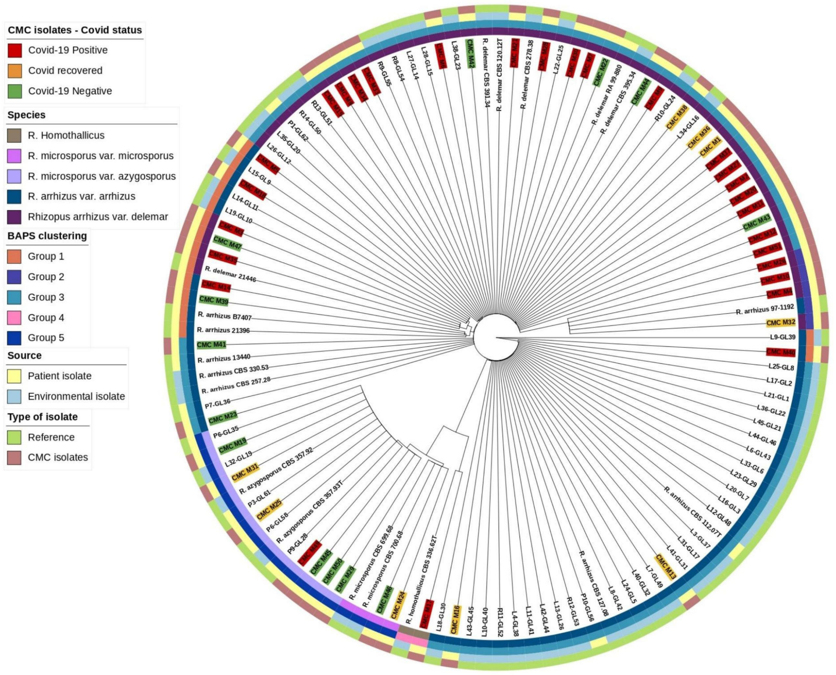


FIGURE 1
Phylogenetic tree based on ITS gene.

followed by Phylogenetic tree construction with RAxML 8.2.11 (Stamatakis, 2014) with the following parameters: GTRGAMMA nucleotide model, Rapid Bootstrapping and search for best-scoring maximum likelihood tree, 100 bootstraps replicates with parsimony random seed 100, followed by visualization and annotation using iTOL (https://itol.embl.de/).

Quality controls

Negative water controls were used for the extraction and subsequent WGS. PhiX controls were used for library preparation. All isolates were blasted with reference isolates. ATCC strains were not used as positive controls.

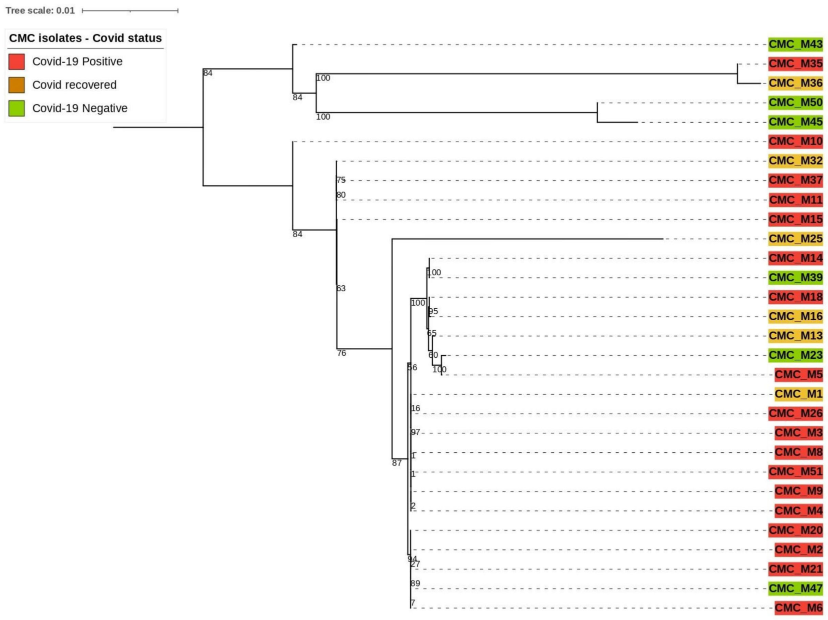


FIGURE 2
Phylogenetic relatedness based on eukaryotic reference markers.

Results

Cohort description

Fifty clinical isolates were cultured from patients diagnosed with ROCM during the study period. The study population included 38 male and 12 female patients with a mean age of 50.28 (28–81) years. Most patients were from Tamil Nadu (34), followed by the adjoining state, Andhra Pradesh (Anand et al., 2022). As shown in Table 1, 28 patients were COVID-19 positive, 13 were COVID-19 negative, and 9 were COVID-19 recovered, and the three groups were compared using ANOVA. All -values <0.05 were considered significant.

The mean duration of symptoms was 10.5 days. A total of 28 patients gave a history of prior hospital admission, and 25 had received steroids. All patients had sinonasal involvement. Forty-six percent had additional orbital involvement, while the palatal and intracranial extension was seen in 36% and 20%, respectively. Extra paranasal sinus involvement was seen predominantly among the COVID-negative patients. Four patients had bony involvement at presentation, while another 14 showed late bony changes.

Among the COVID-19-positive patients, 14 tested positive at admission, while the rest presented within 3 weeks of testing positive. All patients in the cohort had diabetes. Of the seven diabetic patients who presented with ketoacidosis, six were COVID-19 positive. The mean HbA1c was 11.2%, and serum ferritin was 546.8 ng/mL in the cohort. Among the 50 patients with AIFS, 20 patients had associated CGFS.

On follow-up, 40 (80%) patients were alive, among whom 30 had no clinical or radiological evidence of disease, 1 patient had the residual disease, while 9 patients though clinically and endoscopically normal, had radiological changes. Eight (16%) patients had expired, and two were lost to follow-up.

Phylogenetic assessment

Based on the ITS gene (Figure 1), Bayesian analysis of population structure (BAPS) clustering algorithm clustered 50 isolates from CMC and reference isolates into five groups. Group 1 comprises eight of our clinical isolates identified as *R. arrhizus* (*n* = 3) and *R. delemar* (*n* = 5) species closely clustered to the clinical reference strains *R. arrhizus* (B7407 and 21396) and *R. delemar* (21446). Group 2 contains four of our clinical isolates; all were identified as *R. delemar; R. arrhizus* 97-1192 strain shows close relatedness with these groups. Thirty-four clinical isolates were found to be clustered in group 3, 28 CMC isolates clustered with six reference clinical strains, *R. arrhizus* CMC strains (Nagao et al., 2005) were closely related to *R. arrhizus* (13440 and CBS_112.07T), and, similarly, *R. delemar* CMC strains (24) were closely associated with *R. delemar* (RA 99-880). Group 4 consists of one clinical isolate identified as *R. homothallicus* species. Group 5 was divided into two subgroups: 2 clinical isolates identified as *R. microsporus* and 11 clinical isolates (7 isolates from CMC clustered with 4 reference strains) identified as *R. azygosporous*. Table 2 summarizes the BAPS groups stratified by COVID-19-negative, COVID-19-positive, and COVID-19 recovered patients.

TABLE 1 Clinical and laboratory findings of the patients with mucormycosis.

	Total (*N* = 50)	COVID positive (*N* = 28)	COVID recovered (*N* = 9)	COVID negative (*N* = 13)	*p*-value
Age					
Mean ± SD	50.28 ± 12.65	50.50 ± 14.18	48.89 ± 9.78	50.77 ± 12.20	0.938
Range	28-81	28-81	37-61	33-76	
Gender					
Male (%)	38 (76)	19 (67.9)	9 (100.0)	10 (76.9)	0.052
Female (%)	12 (24)	9 (32.1)	0	3 (23.1)	
Comorbidities					
DM	50 (100)	28 (100)	9 (100)	13 (100)	–
DKA	7 (14)	6 (21.4)	1 (11.1)	0	0.077
Duration of symptoms (days)					
Mean ± SD	10.5 ± 25.04	11.21 ± 33.26	11.11 ± 9.03	8.54 ± 7.90	0.950
Range	1–180	1–180	4–28	2–25	
Species					
Rhizopus arrhizus (%)	44 (88)	27 (96.4)	8 (88.9)	9 (69.2)	0.057
Rhizopus microsporus (%)	6 (12)	1 (3.6)	1 (11.1)	4 (30.8)	
Prior hospital admission (%)	28 (56)	17 (60.7)	8 (88.9)	3 (23.1)	0.005

(Continued)

TABLE 1 Continued

	Total (*N* = 50)	COVID positive (*N* = 28)	COVID recovered (*N* = 9)	COVID negative (*N* = 13)	*p*-value
Steroids given (%)	25 (50)	17 (60.7)	6 (66.7)	2 (15.4)	0.010
Disease extent (%)					
Nose and PNS	50 (100)	28 (100)	9 (100)	13 (100)	–
Orbit	23 (46)	13 (46.4)	2 (22.2)	8 (61.5)	0.177
Palate	18 (36)	5 (17.9)	5 (55.6)	8 (61.5)	0.009
Intracranial	10 (20)	5 (17.9)	1 (11.1)	4 (30.8)	0.488
Blood parameters					
HbA1C					0.836
Mean ± SD	11.2 ± 2.35	11.32 ± 2.39	11.34 ± 2.26	11.80 ± 2.54	
Range	5.9–15.6	6.2–>14	8.2–14.9	5.9–>14	
Serum ferritin					
Mean ± SD	546.81 ± 424.64	578.66 ± 446.12	427.86 ± 425.6	567.62 ± 419.24	
Range	120.9–1,909.6	123.1–1,909.6	133.4–1,433.8	120.9–1,423.4	
HPE (%)					0.639
AIFS	29 (58)	16 (57.1)	4 (44.4)	9 (69.2)	
AIFS + CGFS	20 (40)	11 (39.3)	5 (55.6)	4 (30.8)	
No biopsy	1 (2)	1 (3.6)	0	0	
Osteomyelitis (%)					0.082
Nil	32 (64)	18 (64.3)	4 (44.4)	10 (76.9)	
At initial presentation	4 (8)	1 (3.6)	3 (33.3)	0	
Late presentation	14 (28)	9 (32.1)	2 (22.2)	3 (23.1)	
Outcome (%)					0.187
Alive with no clin/radio disease	30 (60)	17 (60.7)	7 (77.8)	6 (46.2)	
Alive with radio+ clinical	9 (18)	4 (14.3)	2 (22.2)	3 (23.1)	
Alive with clinical	1 (2)	1 (3.6)	0	0	
Dead	8 (16)	6 (21.4)	0	2 (15.4)	
LFU	2 (4)	0	0	2 (15.4)	

All tests were compared using ANOVA, $p < 0.05$ was taken as significant.

Based on the mentioned eukaryotic reference marker genes, 30 isolates' genomes were utilized for building the phylogenetic tree to explain the genetic relatedness. This divided 30 clinical CMC isolates into two major branches. The top branch consists of three COVID-19-negative isolates, a COVID-19-positive isolate, and a COVID-19-recovered isolate, and the lower branch consists of 25 isolates composed of 17 COVID-19-positive, 3 COVID-19-negative, and 5 COVID-19-recovered patient isolates.

Phylogenetic tree inference

In our 50 clinical isolates (Figure 1), 31 isolates were identified as *R. delemar*, which includes 22 isolates from COVID-19-positive patients, 4 isolates from COVID-19-recovered patients, and 5 isolates from COVID-19-negative patients; 9 isolates identified as *R.* arrhizus, consisting of 4 COVID-19-positive, 2 COVID-19-recovered, and 3 COVID-19-negative patient isolates; 7 isolates were identified as *R. azygosporous*, comprising 1 COVID-19-positive, 2 COVID-19-recovered, and 4 COVID-19-negative patient isolates; 2 isolates were identified as *R. microsporus* species, which contained one COVID-19-recovered and COVID-19-negative patient; and 1 COVID-19-positive patient isolate belonged to the *R. homothallicus* species. From the BAPS clustering, we observed that 28 of our patient isolates belong to group 3; 17 of them were isolated from COVID-19-positive, 5 from COVID-19-recovered, and 6 from COVID-19-negative patients. All the COVID-19-positive isolates in group 3 were identified as *R. delemar*.

TABLE 2 Distribution of the Mucorales into the various BAPS clusters.

BAP clusters (*n* = 5)	COVID-19-positive patient isolates	COVID-19-recovered patient isolates	COVID-19-negative patient isolates	Total number of patient isolates observed in each group	Species identification by BAPS clustering
Group 1	6	0	2	8	*R. arrhizus* (3), *R. delemar* (5)
Group 2	3	1	0	4	*R. delemar*
Group 3	17	5	6	28	*R. arrhizus strains* 13440 and CBS_112.07T (4), *R. delemar* (28)
Group 4	1	0	0	1	*R. homothallicus*
Group 5	1	3	5	9	*R. microsporus* (2), *R. azygosporous* (7)

BAPS, Bayesian analysis of population structure.

Summary of mutations

Except in CMC_M21 and CMC_M36 isolates, mutations were observed in the CYP51A gene sequences in all 50 isolates with T nucleotide insertion at the 144 and 145 nt, T nucleotide insertion at 263 nt, C330T, C339T, C375T, G798A, A831G, T1008C, and T1542C. In the mutations observed, only the insertion mutations altered the protein sequence, which leads to frameshift since all the other mutations were synonymous. We have also observed mutations in CYP51B at C75T, T129C, insertion of A at the base of 255 nt, C378T, C575T, G1708A, T1144C, and A1485C. In CMC_M1, C575T and G1708A nucleotide changes were observed that lead to protein sequence change P192L and V570I, respectively; all the other mutations observed in this gene sequence of our isolates were synonymous. Insertion mutations were observed in the isolates CMC_M21 and CMC_M36 at CY51B, which lead to the frameshift of protein sequence; these isolates also did not have any mutations at CY51A. Figure 3 summarizes the findings.

Discussion

Mucormycosis is an acute invasive disease-causing rhino-cerebral mucormycosis in patients with diabetes mellitus and immunocompromised patients such as bone marrow or organ transplants. During the delta wave of the COVID-19 pandemic, there were increased cases of mucormycosis among patients with COVID-19 infection. This is the first study looking at the molecular epidemiology of the Mucorales during and after the pandemic.

The molecular epidemiology of mucormycosis has been studied to investigate outbreaks among solid organ transplant patients and compare it with the environmental isolates (Simão et al., 2015). Here in our study, we used WGS to compare the molecular epidemiology and relatedness of the Mucorales in COVID-19-positive, COVID-19-recovered, and COVID-19-negative patients.

As described by our center as well as a study done across India, the most common risk factor for mucormycosis among COVID-19 patients was diabetes mellitus with 21% among the COVID-19-positive patients having diabetes ketoacidosis. The high serum glucose and ferritin levels secondary to uncontrolled diabetes mellitus and ketoacidosis in a hypoxic acidic medium, in combination with COVID-19-induced decreased phagocytosis, stimulated an ideal state for the escalation of mucormycosis. Similar to the case–control investigation done in 11 hospitals across India, the common presentation in our center was rhino-cerebral mucormycosis, both during and after the pandemic (Anand et al., 2022).

Phylogenetic tree

WGS-employed phylogenetic analysis results in a higher resolution in establishing association among isolates. The

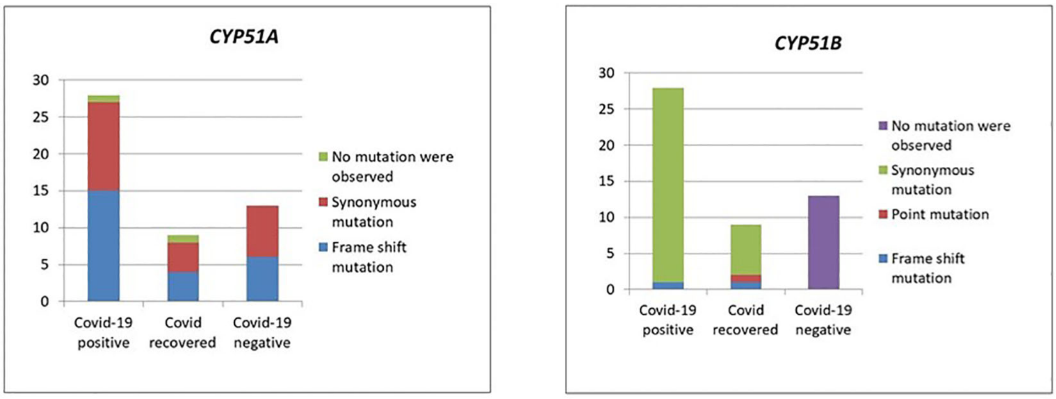


FIGURE 3
Mutations in the CYP51A and CYP51B genes.

phylogenetic tree constructed using ITS1 gene sequences (Figure 1) from the clinical and reference isolates was clustered into five groups. Group 3 consisted of the highest number of clinical isolates, of which COVID-19-positive individuals were found at a higher abundance in the group (Table 2). Groups 3 and 5 showed a dichotomous tree, separated into new groups based on the subspecies type. Using a synoptic approach, clinical isolates were linked to environmental isolates. Three isolates from group 1, group 3, group 4, and group 5 were closely associated with environmental isolates, indicating the involvement of strains from hospital or the surrounding community. In the ITS-based tree, most of the isolates belonged to *R. delemar* species, indicating it to be the responsible strain that had been widely distributed. The top branch homology between the clinical isolates is in the core genome tree. Most of the COVID-19-positive patients were categorized into the *R. delemar* group. A close clustering can also be observed between the COVID-19-positive isolates and the COVID-19-recovered isolates; a close association was also observed between the *R. delemar* and *R. arrhizus* species, indicating a genetic relatedness between them. As observed in several studies, our study finds it tenable that various species are involved in the incidence of mucormycosis outbreaks instead of a solitary strain (Neblett Fanfair et al., 2012; Garcia-Hermoso et al., 2018). Thus, in this study, we were able to ascertain that the *Rhizopus* spp. causing rhino-cerebral mucormycosis during the COVID-19 pandemic were as diverse as the strains after the epidemic. Though we found increased number of group 3 isolates in COVID-19-positive patients, there was no particular clustering of the difference groups, indicating that there was not a common source for the increased surge of cases during the pandemic.

Mutation analysis

Mutation analysis (Figure 3) revealed that in the CYP51A gene, a T nucleotide insertion was observed at the base of 144, 145, and 263 on the sequence that leads to a frameshift of protein translation in 25 isolates (15 COVID-19-positive, 4 COVID-19-recovered, and 6 COVID-19-negative patients).

The following nucleotide changes were also observed along with frameshift mutation in the same isolates: C330T, C339T, C375T, G798A, A831G, T1008C, and T1542C. The A831G, T1008C, C1005T, and C1162T mutations were present without the frameshift mutation in another 23 isolates (from 12 COVID-19-positive, 4 COVID-19-recovered, and 7 COVID-19-negative patients). Since these were synonymous mutations, protein sequences were not altered. Two isolates, CMC_M21 and CMC_M36, from a COVID-positive and a COVID-recovered patient, respectively, did not have any nucleotide changes in the *CYP51A* gene sequence. These two were also the only isolates that had a nucleotide A insertion in the *CYP51B* gene sequence, which led to a frameshift mutation, present along with the following synonymous mutations: C75T, C378T, and T1144C. Like CYP51A, at *CYP51B*, some synonymous mutations, namely, C75T, T129C, C378T, and A1485C, were present without any frameshift mutation in 34 isolates comprising 27 COVID-19-positive and 7 COVID-19-recovered patient isolates, but unlike CYP51A, one COVID-19-recovered patient isolate (CMC_M1) had two point mutations in the CYP51B gene—C575T and G1708A—that led to protein sequence change P192L and V570I, respectively. Of the COVID-19 patients, 13 COVID-19-negative patient isolates did not have any of these mutations in the CYP51B sequence, whereas all isolates had mutations in the CYP51A gene sequence. The CYP51A gene is uniquely responsible for the intrinsic azole resistance phenotype and not CYP51B; CYP51B gene mutations were rarely reported in studies done on fungal isolates. Similarly, we also did not find any mutations in 13 isolates from COVID-negative patients. Further studies are required to understand the CYP51B gene functions and its mutations' involvement in the azole resistance.

Azoles act intracellularly by binding and inhibiting a key enzyme in the ergosterol pathway, lanosterol 14-αdemethylase, a cytochrome P450 enzyme (named ERG11 or CYP51A depending on the fungus) (Jensen, 2016). The mechanism of intrinsic azole intrinsic resistance in Mucorales includes overexpression and/or point mutations in the CYP51A gene. Macedo et al. (2018) analyzed the role of the CYP51A gene and its mutations causing intrinsic resistance to voriconazole and fluconazole in Mucorales. They have demonstrated that the gene sequence of CYP51A can be solely responsible for this intrinsic resistance. They hypothesized that azole resistance in Mucorales would occur because of Y132F and/or F145M substitutions in CYP51A, based on *C. albicans* Erg11p amino acid sequence numbering. In our study, we have not found any point mutation as mentioned above. Limited literature available shows that CYP51 mutations are associated with resistance to voriconazole and fluconazole but not posaconazole or itraconazole (Chau et al., 2006). In addition to synonymous mutations in the CYP51A gene sequence, 25 isolates had a frameshift mutation in the CYP51A gene sequence due to the insertion of T nucleotide at the base of 144, 145, and 263 that leads to alteration in the protein translation. These frameshift mutations were unique and have not been described by Macedo et al. (2018). We have also analyzed the CYP51B gene sequence of our isolates and inferred the results. Based on our observation, we have found that two point mutations, namely, P192L and V570I, in one isolate and one insertion mutation, nucleotide A at the base of 255, led to frameshift of protein translation in two isolates in the case of the CYP51B gene along with some synonymous mutations. Further analysis is required to evaluate the importance of these mutations. The significance of these mutations can be ascertained only by performing *in vitro* and *in vivo* antifungal drug susceptibility testing of these isolates and comparing with the clinical outcome of the patients.

One of the main challenges is that mutations and phylogenetic analysis of Mucorales by WGS has been applied infrequently in studies of mucormycosis because the mucormycete genomes are complex and there are very few scaffolds for assembling the genomes. There are only few clinical isolates in existing databases to compare our isolates for strain relatedness using SNP differences as is being used for other microorganisms.

The main limitation of the study was that the sample size was small and involved those collected over a short period of time during the acute phase of the COVID-19 pandemic and after COVID-19. In non-COVID-19 times, we see mucormycosis in different clinical spectra (Manesh et al., 2019). All patients included in this study during the COVID pandemic had diabetes mellitus as a risk factor. No other immunocompromised status was detected in any of them. Therefore, one of the limitations of the study is that the results of the study may not be representative of patients with other immunocompromised conditions. We also did not perform environmental sampling to look for Mucorales in the community. Thus, we were not able to demonstrate the source of these organisms or demonstrate any particular clustering of clades of *Rhizopus* spp. We also could not compare these genotypic data to phenotypic antifungal susceptibility for mucormycosis.

In summary, WGS on Mucorales is essential to ascertain the phylogenetic relationships among isolates in a hospital or in a community and compare it elsewhere in the country or globally. It also allows for analysis of resistance and virulence markers, which can unravel the biology and pathogenesis of these species. This study emphasizes the need for larger studies to comprehend the molecular epidemiology of these organisms and also the need to standardize WGS-based typing methods for Mucorales and to validate interpretive criteria for strain relatedness.

Data availability statement

The data presented in this study has been deposited and made publicly available in NCBI, https://www.ncbi.nlm.nih.gov/sra/PRJNA993381. Any further enquiries can be directed to the corresponding author.

Ethics statement

This study was approved by the Christian Medical College, Vellore Institutional Review board and Ethics committee (IRB no. 14007).

Author contributions

JM: Design, execution of the study. Analysis of the data, write up and finalisation of the article; MV: WGS and Analysis; MN: Analysis and write up; DS: Collection of Rhizopus isolates, laboratory and clinical data mining; LS: Collection of Rhizopus isolates, laboratory and clinical data mining; LV: Clinical management of patients; RK: Clinical management of patients; RV: Bioinformatics analysis; GC: Bioinformatics analysis. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

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

References

Abe, A., Oda, Y., Asano, K., and Sone, T. (2006). The molecular phylogeny of the genus Rhizopus based on rDNA sequences. *Biosci. Biotechnol. Biochem.* 70 (10), 2387–2393. doi: 10.1271/bbb.60101

Al-Tawfiq, J. A., Alhumaid, S., Alshukairi, A. N., Temsah, M. H., Barry, M., Al Mutair, A., et al. (2021). COVID-19 and mucormycosis superinfection: the perfect storm. *Infection* 49 (5), 833–853. doi: 10.1007/s15010-021-01670-1

Anand, T., Mukherjee, A., Satija, A., Velamuri, P. S., Singh, K., Das, M., et al. (2022). A case control investigation of COVID-19 associated mucormycosis in India. *BMC Infect. Dis.* 22, 856. doi: 10.1186/s12879-022-07844-y

Aranjani, J. M., Manuel, A., Razack, H. I. A., and Mathew, S. T. (2021). COVID-19–associated mucormycosis: Evidence-based critical review of an emerging infection burden during the pandemic's second wave in India. *PloS Negl. Trop. Dis.* 15 (11), e0009921. doi: 10.1371/journal.pntd.0009921

Cendejas-Bueno, E., Kolecka, A., Alastruey-Izquierdo, A., Theelen, B., Groenewald, M., Kostrzewa, M., et al. (2012). Reclassification of the Candida haemulonii complex as Candida haemulonii (C. haemulonii group I), C. duobushaemulonii sp. nov. (C. haemulonii group II) and C. haemulonii var. vulnera var. nov.: three multiresistant human pathogenic yeasts. *J. Clin. Microbiol.* 50 (11), 3641–3651. doi: 10.1128/JCM.02248-12

Chau, A. S., Chen, G., McNicholas, P. M., and Mann, P. A. (2006). Molecular basis for enhanced activity of posaconazole against Absidia corymbifera and Rhizopus oryzae. *Antimicrob. Agents Chemother.* 50 (11), 3917–3919. doi: 10.1128/AAC.00747-06

Cherian, L. M., Varghese, L., Rupa, V., Bright, R. R., Abraham, L., Panicker, R., et al. (2022). Rhino-orbito-cerebral mucormycosis: patient characteristics in pre-COVID-19 and COVID-19 period. *Rhinology* 60 (6), 427–434. doi: 10.4193/Rhin22.099

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., and Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2, e243. doi: 10.7717/peerj.243

Garcia-Hermoso, D., Criscuolo, A., Lee, S. C., Legrand, M., Chaouat, M., Denis, B., et al. (2018). Outbreak of Invasive Wound Mucormycosis in a Burn Unit Due to Multiple Strains of Mucor circinelloides f. circinelloides Resolved by Whole-Genome Sequencing. *mBio* 9 (2), e00573–e00518.

Jensen, R. H. (2016). Resistance in human pathogenic yeasts and filamentous fungi: prevalence, underlying molecular mechanisms and link to the use of antifungals in humans and the environment. *Dan Med. J.* 63 (10), B5288.

Macedo, D., Leonardelli, F., Dudiuk, C., Theill, L., Cabeza, M. S., Gamarra, S., et al. (2018). Molecular confirmation of the linkage between the rhizopus oryzae CYP51A gene coding region and its intrinsic voriconazole and fluconazole resistance. *Antimicrob. Agents Chemother.* 62 (8), e00224–e00218. doi: 10.1128/AAC.00224-18

Manesh, A., Rupali, P., Sullivan, M. O., Mohanraj, P., Rupa, V., George, B., et al. (2019). Mucormycosis—A clinicoepidemiological review of cases over 10 years. *Mycoses* 62 (4), 391–398. doi: 10.1111/myc.12897

Nagao, K., Ota, T., Tanikawa, A., Takae, Y., Mori, T., Udagawa, S.i., et al. (2005). Genetic identification and detection of human pathogenic Rhizopus species, a major mucormycosis agent, by multiplex PCR based on internal transcribed spacer region of rRNA gene. *J. Dermatol. Sci.* 39 (1), 23–31. doi: 10.1016/j.jdermsci.2005.01.010

Neblett Fanfair, R., Benedict, K., Bos, J., Bennett, S. D., Lo, Y. C., Adebanjo, T., et al. (2012). Necrotizing cutaneous mucormycosis after a tornado in Joplin, Missouri, in 2011. *N Engl. J. Med.* 367 (23), 2214–2225. doi: 10.1056/NEJMoa1204781

Nguyen, M. H., Kaul, D., Muto, C., Cheng, S. J., Richter, R. A., Bruno, V. M., et al. (2020). Genetic diversity of clinical and environmental Mucorales isolates obtained from an investigation of mucormycosis cases among solid organ transplant recipients. *Microb. Genomics* 6 (12), mgen000473. doi: 10.1099/mgen.0.000473

Sievers, F., and Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci. Publ Protein Soc* 27 (1), 135–145. doi: 10.1002/pro.3290

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma Oxf Engl.* 31 (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma Oxf Engl.* 30 (9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34 (Web Server issue), W435–W439. doi: 10.1093/nar/gkl200

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 December 2023
DOI 10.3389/fmicb.2023.1225207

Rapid identification of enteric bacteria from whole genome sequences using average nucleotide identity metrics

Rebecca L. Lindsey*, Lori M. Gladney, Andrew D. Huang, Taylor Griswold, Lee S. Katz, Blake A. Dinsmore, Monica S. Im, Zuzana Kucerova, Peyton A. Smith, Charlotte Lane and Heather A. Carleton

Centers for Disease Control and Prevention, Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Atlanta, GA, United States

OPEN ACCESS

EDITED BY
Richard Allen White III,
University of North Carolina at Charlotte,
United States

REVIEWED BY
John Phillip Bowman,
University of Tasmania, Australia
Christopher John Grim,
United States Food and Drug Administration,
United States

*CORRESPONDENCE
Rebecca L. Lindsey
✉ rlindsey1@cdc.gov

RECEIVED 08 August 2023
ACCEPTED 14 November 2023
PUBLISHED 14 December 2023

CITATION
Lindsey RL, Gladney LM, Huang AD, Griswold T, Katz LS, Dinsmore BA, Im MS, Kucerova Z, Smith PA, Lane C and Carleton HA (2023) Rapid identification of enteric bacteria from whole genome sequences using average nucleotide identity metrics.
Front. Microbiol. 14:1225207.
doi: 10.3389/fmicb.2023.1225207

Identification of enteric bacteria species by whole genome sequence (WGS) analysis requires a rapid and an easily standardized approach. We leveraged the principles of average nucleotide identity using MUMmer (ANIm) software, which calculates the percent bases aligned between two bacterial genomes and their corresponding ANI values, to set threshold values for determining species consistent with the conventional identification methods of known species. The performance of species identification was evaluated using two datasets: the Reference Genome Dataset v2 (RGDv2), consisting of 43 enteric genome assemblies representing 32 species, and the Test Genome Dataset (TGDv1), comprising 454 genome assemblies which is designed to represent all species needed to query for identification, as well as rare and closely related species. The RGDv2 contains six *Campylobacter* spp., three *Escherichia/Shigella* spp., one *Grimontia hollisae*, six *Listeria* spp., one *Photobacterium damselae*, two *Salmonella* spp., and thirteen *Vibrio* spp., while the TGDv1 contains 454 enteric bacterial genomes representing 42 different species. The analysis showed that, when a standard minimum of 70% genome bases alignment existed, the ANI threshold values determined for these species were ≥95 for *Escherichia/Shigella* and *Vibrio* species, ≥93% for *Salmonella* species, and ≥92% for *Campylobacter* and *Listeria* species. Using these metrics, the RGDv2 accurately classified all validation strains in TGDv1 at the species level, which is consistent with the classification based on previous gold standard methods.

KEYWORDS

average nucleotide identity, ANI, species identification, enteric bacteria, WGS

Introduction

Conventional bacterial species identification methods, such as phenotypic testing and gene-sequencing analysis, have been employed within the scientific community for years. However, with the increased use of next generation sequencing, new methods are available to analyze the entire DNA of the organisms. This allows for the simultaneous capture of a wide range of information, including whole genes, core genes, and ribosomal genes for species identification and strain typing, characterization of genes for serotype, virulence, antimicrobial resistance, kmer-typing, and much more (Jolley et al., 2012; Besser et al., 2018; Gerner-Smidt et al., 2019a, Gerner-Smidt et al., 2019b; Stevens et al., 2022). More diversity has been identified with sequencing methods than was previously known, due to the limitations of conventional identification methods that rely on shared metabolic characteristics (phenotypic tests) or gene sequencing, which typically only analyze a small fraction of the organism's DNA. This has led to the taxonomic re-classification of entire genera (Yu et al., 2021). The increased use of next generation sequencing also enhances the speed and efficiency of bacterial identification methods, whereas conventional methods were more time-consuming and provided low resolution (Carleton and Gerner-Smidt, 2016; Besser et al., 2018; Gerner-Smidt et al., 2019a; Gerner-Smidt et al., 2019b; Stevens et al., 2022).

Historically, DNA–DNA hybridization (DDH) had been the gold standard for determining prokaryotic species for taxonomic classification (Rossello-Mora and Amann, 2001; Richter and Rossello-Mora, 2009). Rossello discussed the prokaryotic species concept in 2001, "Today, the accepted species classification can only be achieved by the recognition of genomic distances and limits between the closest classified (DNA–DNA similarity), and of those phenotypic traits that are exclusive and serve as diagnostic of the taxon (phenotypic property; Rossello-Mora and Amann, 2001)." This species concept is still applicable today; however, the genomic comparisons are now based on whole genome sequence (WGS) analysis.

In 2005, the average nucleotide identity (ANI) method was shown to be a plausible substitute for DDH since a 70% DDH threshold for species classification correlated well with a 94% ANI similarity threshold. This method, proposed by Kostantinidis et al., used pairwise alignment (BLAST) to identify the best hits of shared orthologous gene content between genomes being compared, obtaining the ANIb values (Konstantinidis and Tiedje, 2004; Goris et al., 2007; Richter and Rossello-Mora, 2009; Rodriguez-R, 2016). However, a drawback of ANIb is the need to perform gene prediction on the assembly before an ANI score can be determined.

Later methods eliminated the need for this prediction step by using local alignments of sequences of varying length and similarity. In 2007, Goris et al. expanded on the ANIb method by generating 1,020 bp fragments of the query genome and compared the ANI between the fragments and a reference genome using BLAST (Goris et al., 2007). In 2009, Richter et al. implemented an ultra-fast alignment tool, which compared the entire WGS contigs between genomes using the nucmer alignment program in MuMMer software, to calculate ANI values, referred to as ANIm (Kurtz et al., 2004). Kurtz et al. provided a dnadiff wrapper, which compares the resulting output files from the nucmer alignment program, to simplify and summarize ANIm output metrics for the differences between two genomes (Kurtz et al., 2004). Jain et al. further developed ANI methods by implementing FastANI, which is a method based on the minHash algorithm and read mapping. FastANI, similar to ANIb, aims to identify reciprocal or orthologous mappings and has an 80% identity cutoff (Ondov et al., 2016; Jain et al., 2018). FastANI has shown results that are comparable to the previous methods but has significantly improved the overall runtime to just seconds (Jain et al., 2018). GAMBIT was recently described as a kmer-based method comparable in accuracy and speed to FastANI (Lumpe et al., 2023). GAMBIT computes Jaccard distances based on a subset of the genome's kmers and, similar to FastANI, uses raw sequencing reads (Lumpe et al., 2023).

Additional methods for species and subspecies identification have also been described. Ribosomal MLST was described by Jolley et al. (2012), but this method requires gene prediction, unlike ANIm and FastANI (Jolley et al., 2012). More recently, a new method for ribosomal MLST nucleotide identity (r-MLST-NI) has been developed for classifying *Klebsiella* and *Raoultella* species and may be useful for identifying other bacterial species (Bray et al., 2022). Public health laboratories in the United States, including our laboratory, have transitioned to WGS analysis from conventional methods for identification and surveillance of enteric pathogens. For this transition, a rapid and an easily standardized method of species identification using WGS was needed, which could be easily integrated into the PulseNet national molecular surveillance system [National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), 2021] for enteric pathogens. In this study, we describe the implementation of an accurate, rapid, stand-alone, sequence-based method for the identification of *Campylobacter*, *Escherichia/Shigella*, *Listeria*, *Salmonella*, and *Vibrionaceae* species. This method is comparable to previous gold standard methods and utilizes the ANIm method. We compared over 450 genome assemblies to set the threshold ANIm values consistent with conventional identification methods. This method is currently employed for the precise speciation of enteric organisms from WGS using the Reference Genome Dataset version 2 (RGDv2) in BioNumerics and on the command-line, for routine identification of *Campylobacter*, *Escherichia/Shigella*, *Listeria*, *Salmonella*, and *Vibrionaceae* species.

Materials and methods

Genome selection for ANI detection

For this study, we selected two sets of genomes which included the Reference Genome Dataset version 2 (RGDv2, Supplementary Table 1) and the Test Genome Dataset version 1 (TGDv1, Supplementary Table 2). The strains were selected from genome assemblies available on NCBI or from the PulseNet Reference Outbreak Surveillance Team's (PROST) enteric bacteria reference collections to represent the diversity of enteric bacteria. These well-characterized strains were previously identified by methods, such as phenotypic characterization, gene sequencing, phylogenetic analysis of the *rpoB* gene, and Accuprobe® (*Listeria monocytogenes*). All sequences met the standard PulseNet QAQC metrics, including a minimum Q score of 30, and sequencing coverages for downstream analysis: 40× for *Escherichia*, *Vibrio*, and *Shigella*, 30× for *Salmonella* and *Campylobacter*, and 20× for *Listeria* (Tolar et al., 2019).

The RGDv2 (Supplementary Table 1) included all species characterized as part of PulseNet, and the set was minimized for rapid analysis. It comprised 43 genome assemblies representing 32 enteric species, consisting of 10 assemblies representing 6 *Campylobacter* spp., 3 assemblies representing 3 *Escherichia/Shigella* spp., 11 assemblies representing 6 *Listeria* spp., 2 *Salmonella* assemblies representing 2 species, and 15 *Vibrionaceae* assemblies representing 11 *Vibrio* species, 1 *Grimontia* species, and 1 *Photobacterium* species. The RGDv2 assemblies were sequenced by Illumina, PacBio, or both instruments. The WGS reads for RGDv2 references were assembled using SPAdes for Illumina data (Bankevich et al., 2012) and HGAP (University M, 2014) for PacBio data. *Escherichia* and *Vibrio* genomes are larger and more complex due to phage regions; these assemblies were generated using both Illumina and PacBio sequencers. The NCBI BioSample data include additional information regarding sequencing chemistry and assembly methods for all strains.

The TGDv1 consists of 454 genome assemblies from 42 different enteric bacterial species (Supplementary Table 2), including the RGDv2 genome assemblies, and it is designed to represent all species necessary for querying identification, as well as rare and closely related species, to confirm the accuracy of ANIm for correct identification of species. The TGDv1 genomes were assembled using SPAdes v3.11 with default options (Bankevich et al., 2012).

Development of custom ANI scripts

We developed custom scripts to utilize the dnadiff workflow in MUMmer v3.23 (Kurtz et al., 2004), facilitating pairwise comparisons with references and generating results in a tabular format. These scripts were developed for the command line. These scripts are published on our GitHub site (NCEZID-biome, 2021). The ANIm script runs on dnadiff and parses the field "AvgIdentity" to detect the percent identity. Additionally, to measure the breadth of the alignment, the script parses the AlignedBases field. To ensure consistency, the same ANIm script runs on both the command line (ani-m.pl) and as a plugin for BioNumerics (ani-m-bionumerics.pl).

Determination of ANI metrics

The TGDv1 genomes were supplied as the reference and the query; the genomes were compared in a pairwise, all-vs-all fashion. The RGDv2 genomes, our gold standard set of references, were included in TGDv1 and the threshold analysis.

We used the ggplot2 and dplyr modules in R to analyze and generate a scatter plot of the values for ANI and percent aligned bases for all comparisons. Additionally, a violin plot was created from the ANI values for a subset of species represented in RGDv2. For the violin plot, only ANI comparisons with a minimum of 70% aligned bases were examined to ensure that percent ANI was being calculated over significant portions of the genome and to avoid spurious high percent ANI matches over repetitive regions.

Down sampling for limits of detection

The reads for representative species of RGDv2 including two *Campylobacter*, three *Escherichia*, one *Listeria*, two *Salmonella*, and three *Vibrio* were downsampled to various coverage levels: 0.5×, 1×, 5×, 10×, 15×, 20×, 30×, 40×, and 50×. A 1× coverage was calculated as the total assembly size of the original coverage SPAdes assembly. The desired coverage and the total number of bases in the raw reads were used to calculate a percentage of the reads needed for that coverage level. Subsequently, we used the Fasten package (lskatz, 2023) to sample enough reads to meet the expected coverage. The coverage level was verified using the read metrics script in CG-Pipeline (Kislyuk et al., 2010). These downsampled reads were used to assemble each genome as previously described in this study. Most genomes at 0.5× and 1× could not be assembled with SPAdes and could not be used as assemblies for the 0.5× and 1× coverage level analyses.

At each downsampling level of every genome, we recorded the N50, a standard assembly metric. Then, we computed the ANIm method against the reference genome for each coverage level. We noted the change in the ANI value received at the different coverage levels as compared to the 50× downsampled assembly.

Comparison of ANI methods: time trials and method compatibility

Pairwise ANI comparisons were generated using TGDv1 genomes, which were run in an all-vs-all fashion using the ANIm, FastANI, and ANIb algorithms, to evaluate the amount of time each method took from the launch of the script to report of the result. This workflow is encoded on our GitHub site (NCEZID-biome, 2021) as the *launch_all_ani* shell script. For each algorithm, we computed the ANI value and recorded the duration of each analysis using GNU time. Pairwise scatterplots for each pair of algorithms were plotted using ANI results, and a trend line was computed in Microsoft Excel; only algorithm pairs involving ANIm were included. Additionally, the frequency of the analysis durations for each algorithm were computed and plotted in Microsoft Excel.

Results

Determination of ANI metrics

Computing the ANI of a query genome against a reference genome yields both the ANI value and the percentage of bases aligned. The percent bases aligned metric conveys what percentage of the reference genome is shared with the query. In this study, we compared the 454 TGDv1 genome assemblies in an all-vs-all comparison using ANIm (Supplementary Table 2), which resulted in 206,116 total comparisons. We plotted the percent bases aligned against the ANI for all genera and color-coded the between-species and within-species values (Figure 1). We noted that all the within-species ANI values appeared when the percent bases aligned was above 70%, consistent with our percent bases aligned threshold for excluding spurious high ANI matches.

By plotting all-vs-all ANI, we observed that the ANIm method effectively distinguished within-species comparisons from between-species comparisons, enabling the determination of thresholds for relevant species (Figure 2). The ANI threshold values were ≥ 95% for *Escherichia*/*Shigella* and *Vibrionaceae* species, ≥93% for *Salmonella* species, and ≥ 92% for *Campylobacter* and *Listeria* species; the ANIm method accurately classified all validation strains in the TGDv1 at the species level, when considering comparisons across >70% of bases aligned (Table 1). In this study, we identify an ANI threshold for each genus as shown in Table 1 based on the results of the ANIm analysis. Notably, *Vibrionaceae* and *Escherichia* species have a 95% threshold, while species from *Salmonella*, *Campylobacter*, and *Listeria* have a lower ANI threshold for distinguishing within-species from between-species comparisons (92–93%) when a ≥70% alignment threshold is met. We used traditional taxonomic definitions of these species that rely on phenotypic tests to verify these within-species and between-species comparisons (Ciufo et al., 2018). Some of these lower ANI thresholds may be the attributed to the greater diversity that

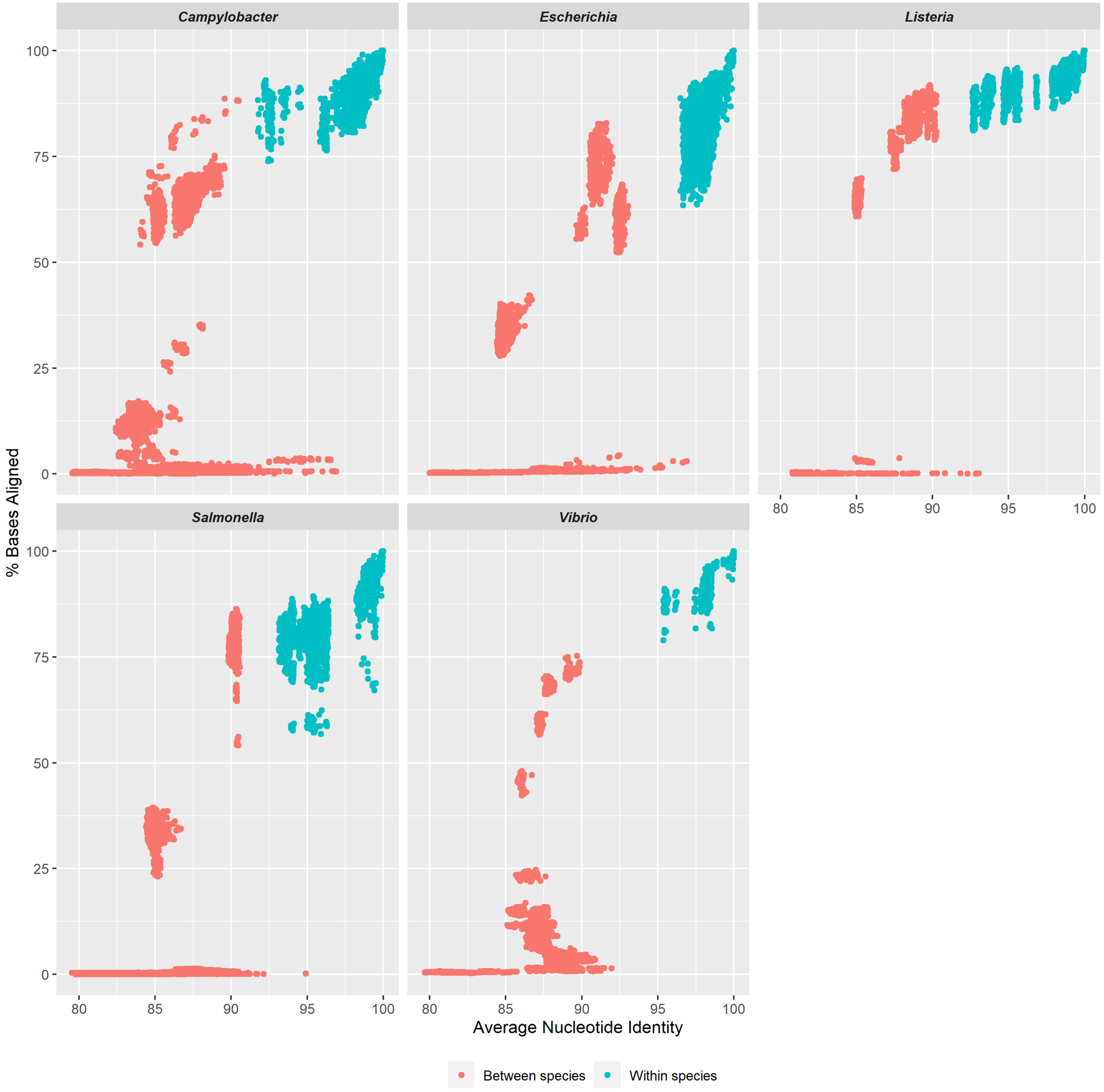


FIGURE 1
ANI limits for enteric detection. Scatter plots of average nucleotide identity versus percent aligned bases for four genera and one family: *Campylobacter*, *Escherichia*, *Listeria*, *Salmonella*, and *Vibrionaceae*. Each plot displays the relationship between ANI and percent aligned bases (e.g., reference genome alignment coverage) for both within-species and between-species in each group.

WGS-based methods can capture compared to the conventional naming of *Salmonella*, *Campylobacter*, and *Listeria* species.

Down sampling for limits of detection

To determine the robustness of the ANIm method at different coverage levels, an experiment was conducted to determine the lowest depth of coverage of a genome assembly required for accurate species identification. Several assemblies from representative species were assembled from coverage depths of 50× to 0.5× to find where an ANI value starts deviating (Figure 3). After down sampling, most genomes at 0.5× and 1× could not be assembled with SPAdes. In some cases, identification was made at 5× coverage, especially for *Salmonella* and *Listeria* genomes. For all enteric species in RGDv2, we determined a minimum of 10× depth-of-coverage for genome assemblies. In the standard bioinformatic analysis for molecular surveillance within PulseNet, the sequencing depth cutoffs are 40× for *Escherichia*, *Vibrionaceae* and *Shigella*, 30× for *Salmonella* and *Campylobacter*, and 20× for *Listeria*, which makes ANIm compatible with this public health usage (Tolar et al., 2019).

Comparison of ANI methods: time trials and method compatibility

We compared several methods to calculate ANI: ANIb, ANIm, and FastANI. We first compared these three methods in a speed trial (Figure 4), examining the range of ANI runtimes for pairwise comparisons. An all-*vs*-all comparison of the TGDv1 showed that FastANI trials produced the fastest results, followed by ANIm and ANIb. Peak frequency runtimes for FastANI (approximately 0.75 and

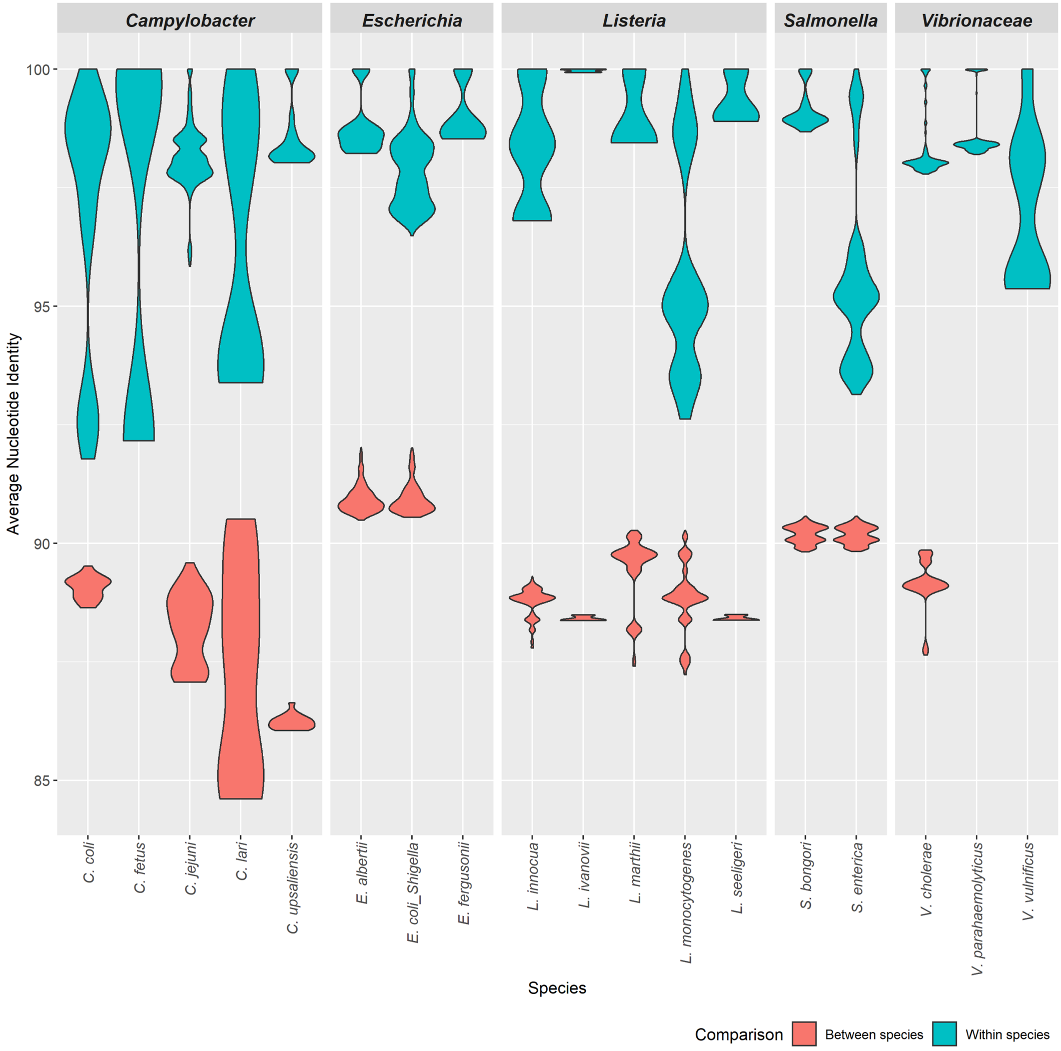


FIGURE 2
ANI values for five genera. Violin plots show ANI ranges for five genera: *Campylobacter, Escherichia, Listeria, Salmonella* and *Vibrio*. Each plot displays the variation in ANI values for both within a species (blue) and between a species (red) in each group.

TABLE 1 Taxon-specific values for identification by ANI.

Taxon	ANI value (%)	Aligned bases (%)	Genome size (Mb.)
Campylobacter spp.	≥92	≥70	1.4 to 2.2
Escherichia spp.	≥95	≥70	4.5 to 5.5
Listeria spp.	≥92	≥70	2.7 to 3.1
Salmonella spp.[1]	≥93	≥70	4.56 to 5.5
Vibrionaceae spp.	≥95	≥70	3.8 to 6.2

Species level identification results are reported for query assemblies with ANI values listed below for *Campylobacter*, *Escherichia*, *Listeria*, *Salmonella*, and *Vibrionaceae* species. Taxon, ANI value (% value for ANI lower cutoff), aligned bases (%) and genome size (in megabases) for each species are listed. 1ANI can be used to identify one clinically important subspecies, *Salmonella enterica* subspecies enterica when the ANI score against the *Salmonella enterica* reference is >98%. Individual species thresholds may ultimately differ for *Salmonella bongori*, as all isolates tested to date result in >98% ANI score, >85% coverage, and lengths up to 5.0 Mb.

2 s), ANIm (approximately 2 and 4 s), and ANIb (approximately 9 s) were observed; two different frequency peaks were noted for ANIm and FastANI. FastANI, while being an order of magnitude faster than ANIm, lacks an alignment report that includes the number or percentage of aligned bases, similar to ANIb. We selected ANIm as a preferred method due to speed, and it has provided the desired output of ANI score and percent genome alignment.

Using the same results from the time trials, we next measured the similarity between the results when comparing FastANI to ANIm and ANIb to ANIm (Figure 5). We plotted the percent identity of ANIb and FastANI against ANIm to form a scatterplot. This benchmark shows a trendline with FastANI: $y = 1.2376\times - 23.245$ ($R^2 = 0.9741$) and ANIb: $y = 1.463\times - 45.49$ ($R^2 = 0.9124$). The R^2 scores suggest a correlation between ANIb, ANIm, and FastANI. However, ANIb and FastANI often reported ANI scores of 0, a null value, when compared against distantly related species; instances of null ANI scores were excluded in our benchmark analysis. ANIb and FastANI do not consider low identity regions in their calculations, and ANIb and FastANI report these null ANI scores when the scores fall below 60 and 80%, respectively (Konstantinidis and Tiedje, 2005; Jain et al., 2018). Alternatively, ANIm does not have this requirement and null ANI values were never reported for ANIm.

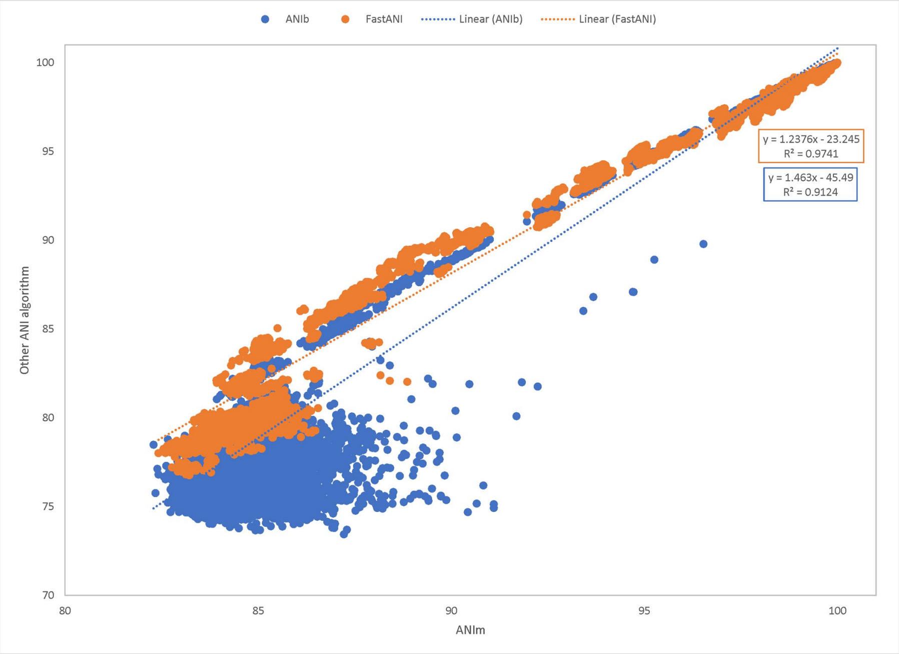


FIGURE 3
Downsampling for limits of detection. Representative species of *Campylobacter*, *Escherichia*, Listeria, *Salmonella*, and *Vibrio* were downsampled from 50× to 0.5× and analyzed with the ANIm algorithm. Genome coverage is plotted on the x-axis; the natural log of N50 (lnN50) is plotted on the left y-axis; and percent change from ANI at 50× is plotted on the right y-axis. The dotted blue line shows the average N50 for all the assemblies. The dark green line indicates the aggregate ANI values, or the average percentage that each ANI value deviated from what it was at 50×. Coverage cutoff of 10× was established based on this analysis, as species identification is not reliable below 10×. Additionally, the aggregate ANI begins accruing below 10×, gaining larger standard deviations.

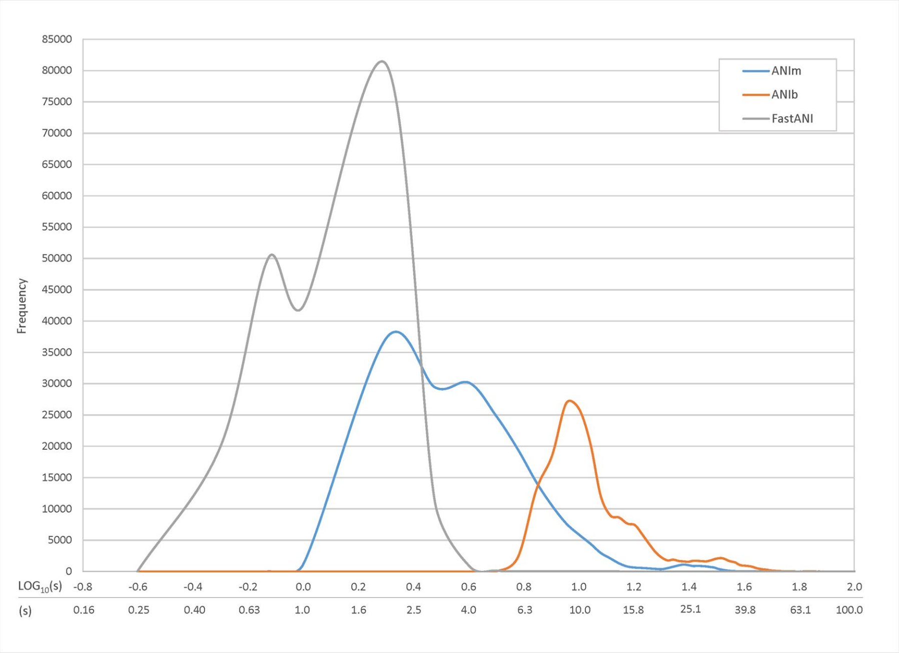


FIGURE 4
Individual Query Speed by ANI Method. Time trials were conducted to compare the runtime of three different ANI methods: ANIb, ANIm, and FastANI. TGDv1 genomes were compared against each other, and 206,116 total comparisons were generated along with their associated runtimes. Approximately 0.10% (ANIm) and 0.02% (ANIb) of the comparisons were excluded because they exceeded the maximum graphical runtime of 100 s; there were no comparisons excluded for FastANI. The most common runtimes were approximately 9 s for ANIb, 2 and 4 s for ANIm, and 0.75 and 2 s for FastANI; two different frequency peaks were noted for ANIm and FastANI.

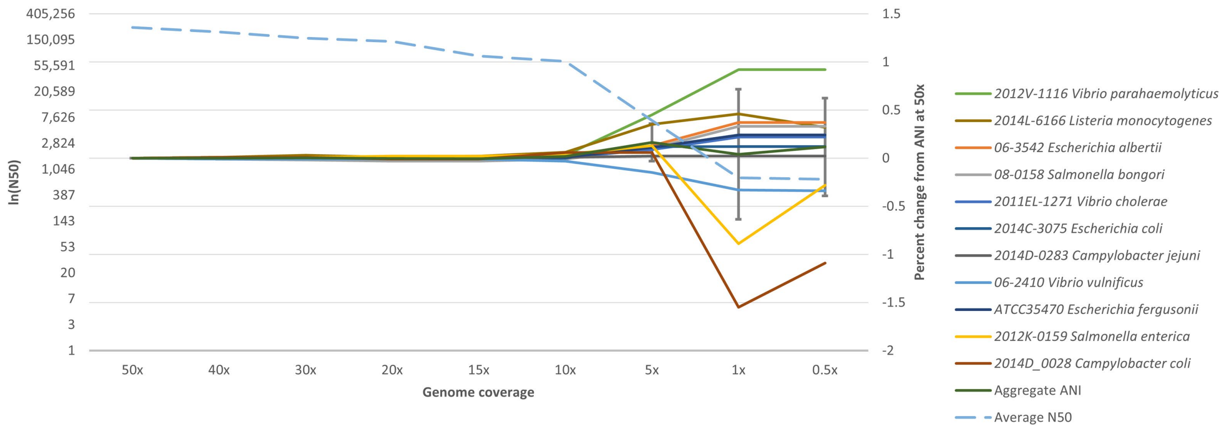


FIGURE 5
Pairwise comparisons of ANIb and FastANI to ANIm. ANIm is plotted on the x-axis while ANIb and FastANI are plotted on the y-axis. All data satisfied the ANIm metric of greater than 70% aligned bases. A goodness-of-fit was detected for each method. FastANI's slope is close to one (FastANI: y = 1.2376–23.245with an R^2 = 0.9741), while ANIb's slope is also close to one (ANIb: y = 1.463x – 45.49 with an R^2 = 0.9124).

When removing null percentages, ANIb scores ranged from 73.43 to 100.00 with Q1, median, and Q3 being 77.01, 79.55, and 89.00, respectively (Supplemental Table 3). Similarly, FastANI scores ranged from 76.76 to 100.00 with a median of 82.15, Q1 of 81.75, and Q3 of 95.11. Similarly, the associated ANIm scores ranged from 82.42 to 100.00 with a median of 84.98 (Q1 and Q3: 84.47 and 95.23) for the FastANI trendline and ANIm scores from 82.29 to 100.00 with a median of 85.15 (Q1 and Q3: 84.45 and 90.21) for the ANIb trendline. Inclusion of additional ANIm scores, which were associated with null percentages in either ANIb or FastANI, had an adjusted range of 78.51100.00 with a median of 83.48, Q1 of 81.53, and Q3 of 85.6 (Supplemental Table 3).

An outline of the ANIm species identification method is illustrated in Figure 6. For routine identification, ANI values are calculated for genome assemblies that meet or exceed the alignment criteria of 70% aligned bases with an RGDv2 reference(s). If the threshold meets the cutoffs per species (Table 1), then a species identification is reported.

Discussion

The ANIm method described here allows for rapid, quantitative, and accurate species identification using the WGS data from enteric bacteria. We have implemented an ANIm methodology on the UNIX command line and in BioNumerics version 7.6 for routine identification of *Campylobacter*, *Escherichia/Shigella*, *Listeria*, *Salmonella*, and *Vibrionaceae* species. The ANIm value and percent bases aligned describe the extent to which one genome assembly is identical to another and can be used to determine the species identity of an assembled query genome by comparing it to a database of reference genomes with historically described taxonomy. To generate this reference genome database for ANIm species identification, we assembled the RGDv2, which contains 43 high-quality representative genomes for relevant PulseNet species, whose species identity had been established with previous gold standard methods (Supplementary Table 1). Any genome assembly can be compared against the reference genomes found in the RGDv2 for species identification. This smaller representative set of reference genomes was chosen to make this identification faster. To expand ANI speciation to other species, a representative genome or genomes of the species of interest, after validation, can be added to the RGDv2 (Supplementary Table 1).

We determined the thresholds for species identification with ANIm by comparing the enteric bacterial genomes from TGDv1, which comprised 454 genomes, including the RGDv2 genomes, whose species identity had also been previously established using gold standard methods. The analysis showed that ANI threshold values of ≥95% for *Escherichia/Shigella* and *Vibrionaceae* species, ≥93% for *Salmonella* species, and ≥ 92% for *Campylobacter* and *Listeria* species classified all validation strains in TGDv1 accurately at the species level, when considering comparisons across >70% of bases aligned. The ANIm thresholds reported in this study are similar to the previously published species boundaries for ANIb (94%), ANIm (95–96%), and FastANI (95%; Konstantinidis and Tiedje, 2005; Richter and Rossello-Mora, 2009; Jain et al., 2018). The lower ANI boundaries (92–93%) observed in this study for *Salmonella*, *Campylobacter*, and *Listeria* may be due to a wider degree of diversity within the species of those genera. As new species may be identified for these genera, we will re-evaluate our ANI thresholds. Moreover, we performed downsampling experiments to examine how genome coverage levels affect the ability of the ANIm tool to provide a result consistent with gold standard methods, and we found that reliable speciation using ANIm can be achieved with genomes assembled from ≥ sequencing read coverage of 10× or greater.

We compared three different methods for computing ANI: ANI using BLAST (ANIb), ANI using MuMMer (ANIm), and FastANI. We focused our comparison on these ANI methods and evaluated them for speed, accuracy, and easy interpretation. While all three of the ANI methods tested were comparable in speed and accuracy, ANIm was the easiest to standardize and interpret using the ANI and percent bases aligned metrics provided by the dnadiff wrapper script. We compared ANIm to ANIb and FastANI by correlating the ANI values from pairwise comparisons across the

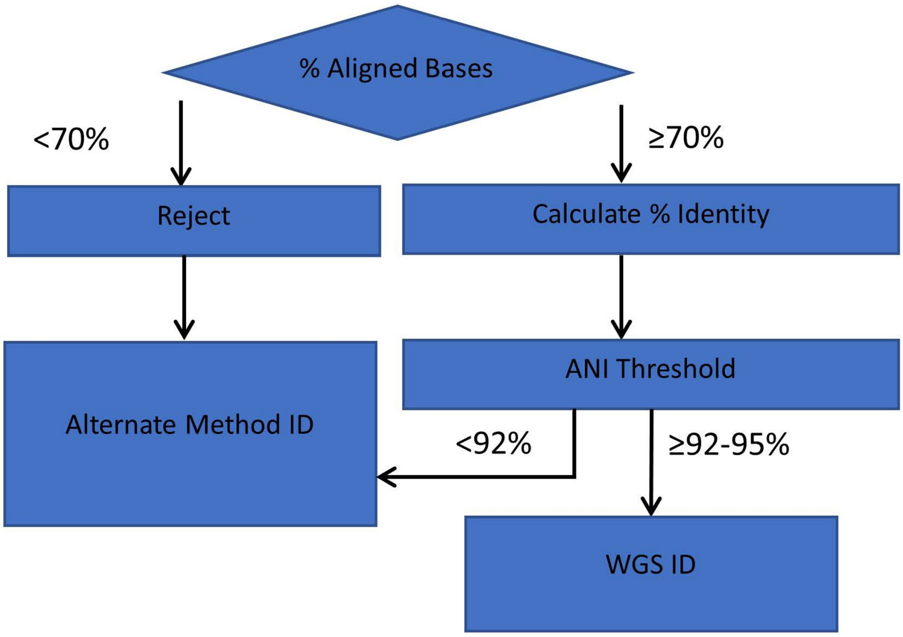


FIGURE 6
Workflow diagram for ANI method of identification. Genome comparisons with >70% of aligned bases and lower ANI Thresholds (≥92–95%, depending on species) are acceptable for interpretation and reporting with the ANI identification workflow.

TGDv1 genome set. All three methods produced comparable ANI results with correlation coefficients of 1.24 and 1.46 and high R^2 scores (>0.9), for both the correlation of FastANI to ANIm and ANIb to ANIm. Additionally, we evaluated the differences in speed of the three distinct tools. All three of the ANI methods had median run times of less than 10 s for a pairwise comparison. To the best of our knowledge, this is the first comparison of the runtime for ANIm and FastANI. FastANI analyses were generally completed faster than ANIm and ANIb, and ANIm was somewhere in the middle from job submission to result. However, overlap was observed in runtimes among all three tools. As all tools demonstrate efficient performance within the range of 10 s or less, the variations in runtimes are likely not significant until a large number of comparisons are being analyzed. While other methods, such as ribosomal MLST, ribosomal MLST nucleotide identity (r-MLST-NI), and k-mer based methods like GAMBIT, hold promise for bacterial species identification, it is important that these methods were not evaluated in this study.

In this study, we have implemented ANI for enteric species identification using MUMmer (ANIm) and demonstrated the utility of ANI for species identification. Furthermore, we simplified ANI-based enteric species identification using a new standard database, RGDv2, built from reference genomes identified with previous gold standard methods and demonstrated its robustness. We also showed that only 10× sequencing coverage is needed to reliably detect species using RDGv2. This low coverage requirement and the speed of the ANIm analysis are advantageous when turnaround time is crucial, as is common in public health settings. For further variant analysis, we have higher coverage requirements in PulseNet. An opportunity for future development may include evaluating the robustness of ANI with additional genome assembly methods compatible with both short-and long-read sequencing methods. The approach here is also generalizable for any situation, where a set of organisms need to be rapidly identified for species by adding and validating reference species genomes to an ANIm database.

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

LG, TG, AH, LK, and RL: conceptualization, validation, visualization, and writing—review and editing. LG, AH, and LK: software. TG: data curation. LG, AH, LK, and TG: formal analysis. CL, PS, MI, BD, and ZK: investigation. RL: project administration and writing original draft. HC: final review.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was made possible through support from the Advanced Molecular Detection (AMD) Initiative grant number AMD-21 at the Centers for Disease Control and Prevention. Apart from the funding, the Office of AMD did not have any further involvement in the study. This project was supported, in part, by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

Acknowledgments

The authors wish to thank Cheryl Tarr for the original idea of utilizing ANI metrics for species identification.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1225207/full#supplementary-material

References

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

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., and Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* 24, 335–341. doi: 10.1016/j.cmi.2017.10.013

Bray, J. E., Correia, A., Varga, M., Jolley, K. A., Maiden, M. C. J., and Rodrigues, C. M. C. (2022). Ribosomal MLST nucleotide identity (rMLST-NI), a rapid bacterial species identification method: application to Klebsiella and Raoultella genomic species validation. *Microb. Genom.* 8, 1–14. doi: 10.1099/mgen.0.000849

Carleton, HA, and Gerner-Smidt, P. (2016). Public health microbiology is undergoing its biggest change in a generation, replacing traditional methods with whole-genome sequencing. Microbe Magazine. 311–317.

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., et al. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* 68, 2386–2392. doi: 10.1099/ijsem.0.002809

Gerner-Smidt, P., Besser, J., Concepcion-Acevedo, J., Folster, J. P., Huffman, J., Joseph, L. A., et al. (2019a). Whole genome sequencing: bridging one-health surveillance of foodborne diseases. *Front. Public Health* 7:172. doi: 10.3389/fpubh.2019.00172

Gerner-Smidt, P., Besser, J., Concepcion-Acevedo, J., Folster, J. P., Huffman, J., Joseph, L. A., et al. (2019b). Corrigendum: whole genome sequencing: bridging one-health surveillance of foodborne diseases. *Front. Public Health* 7:365. doi: 10.3389/fpubh.2019.00365

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

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

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 158, 1005–1015. doi: 10.1099/mic.0.055459-0

Kislyuk, A. O., Katz, L. S., Agrawal, S., Hagen, M. S., Conley, A. B., Jayaraman, P., et al. (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26, 1819–1826. doi: 10.1093/bioinformatics/btq284

Konstantinidis, K. T., and Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3160–3165. doi: 10.1073/pnas.0308653100

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12. doi: 10.1186/gb-2004-5-2-r12

lskatz (2023). fasten. Available at: https://github.com/lskatz/fasten (Accessed May 18, 2023).

Lumpe, J., Gumbleton, L., Gorzalski, A., Libuit, K., Varghese, V., Lloyd, T., et al. (2023). GAMBIT (genomic approximation method for bacterial identification and tracking): a methodology to rapidly leverage whole genome sequencing of bacterial isolates for clinical identification. *PLoS One* 18:e0277575. doi: 10.1371/journal.pone.0277575

National Center for Emerging and Zoonotic Infectious Diseases (NCEZID) (2021). DoF, waterborne, and environmental diseases (DFWED). Pulse Net. Available at: https://www.cdc.gov/pulsenet/ (Accessed May 18, 2023).

NCEZID-biome (2021). ANI-paper. Available at: https://github.com/ncezid-biome/ANI-paper (Accessed May 18, 2023).

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using min hash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Richter, M., and Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Rodriguez-R, L. M. K. K. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ* 4:e1900v1. doi: 10.7287/peerj.preprints.1900v1

Rossello-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67. doi: 10.1016/S0168-6445(00)00040-1

Stevens, E. L., Carleton, H. A., Beal, J., Tillman, G. E., Lindsey, R. L., Lauer, A. C., et al. (2022). Use of whole genome sequencing by the Federal Interagency Collaboration for genomics for food and feed safety in the United States. *J. Food Prot.* 85, 755–772. doi: 10.4315/JFP-21-437

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., et al. (2019). An overview of pulse net USA databases. *Foodborne Pathog. Dis.* 16, 457–462. doi: 10.1089/fpd.2019.2637

University M (2014). Pac Bio HGAP genome assembly pipeline. Available at: https://jtremblay.github.io/pipelines/2014/05/05/Pac Bio-HGAP3-pipeline. (Accessed May 18, 2023).

Yu, D., Banting, G., and Neumann, N. F. (2021). A review of the taxonomy, genetics, and biology of the genus Escherichia and the type species *Escherichia coli*. *Can. J. Microbiol.* 67, 553–571. doi: 10.1139/cjm-2020-0508

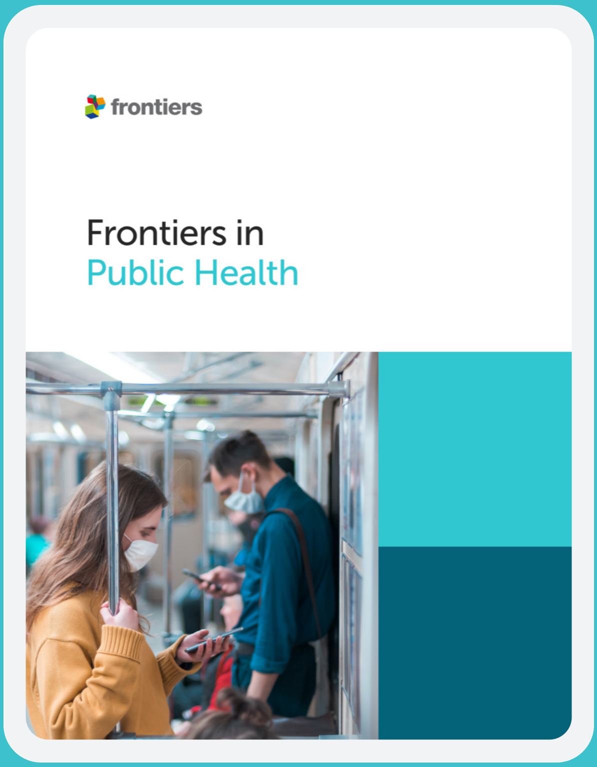
Frontiers in
Public Health

Explores and addresses today's fast-moving healthcare challenges

One of the most cited journals in its field, which promotes discussion around inter-sectoral public health challenges spanning health promotion to climate change, transportation, environmental change and even species diversity.

Discover the latest Research Topics

See more →


frontiers

Frontiers in
Public Health

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact