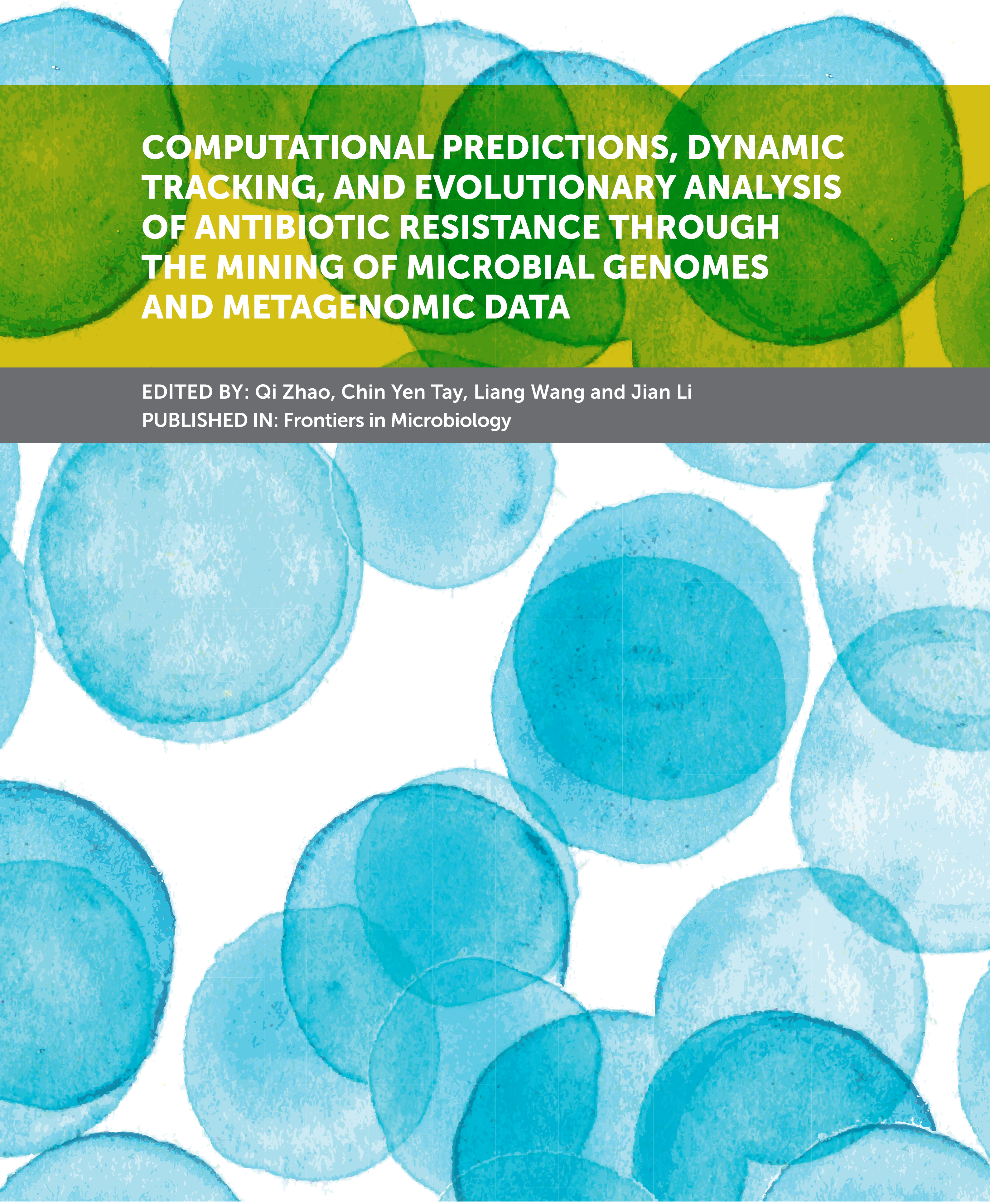


COMPUTATIONAL PREDICTIONS, DYNAMIC TRACKING, AND EVOLUTIONARY ANALYSIS OF ANTIBIOTIC RESISTANCE THROUGH THE MINING OF MICROBIAL GENOMES AND METAGENOMIC DATA

EDITED BY: Qi Zhao, Chin Yen Tay, Liang Wang and Jian Li

PUBLISHED IN: Frontiers in Microbiology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-013-8
DOI 10.3389/978-2-88976-013-8

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

COMPUTATIONAL PREDICTIONS, DYNAMIC TRACKING, AND EVOLUTIONARY ANALYSIS OF ANTIBIOTIC RESISTANCE THROUGH THE MINING OF MICROBIAL GENOMES AND METAGENOMIC DATA

Topic Editors:
Qi Zhao, University of Science and Technology Liaoning, China
Chin Yen Tay, University of Western Australia, Australia
Liang Wang, Xuzhou Medical University, China
Jian Li, Tulane University, United States

Citation: Zhao, Q., Tay, C. Y., Wang, L., Li, J., eds. (2022). Computational Predictions, Dynamic Tracking, and Evolutionary Analysis of Antibiotic Resistance Through the Mining of Microbial Genomes and Metagenomic Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-013-8

Table of Contents

frontiers | Frontiers in Microbiology

EDITORIAL
published: 04 April 2022
doi: 10.3389/fmicb.2022.880967

Editorial: Computational Predictions, Dynamic Tracking, and Evolutionary Analysis of Antibiotic Resistance Through the Mining of Microbial Genomes and Metagenomic Data

Liang Wang[1,2,3], *Alfred Chin Yen Tay*[4], *Jian Li*[5] *and Qi Zhao*[6]*

[1] Department of Bioinformatics, School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou, China, [2] Jiangsu Key Laboratory of New Drug Research and Clinical Pharmacy, School of Pharmacy, Xuzhou Medical University, Xuzhou, China, [3] Laboratory Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China, [4] The Marshall Centre for Infectious Diseases, Research and Training, University of Western Australia, Perth, WA, Australia, [5] School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States, [6] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China

Keywords: antibiotic resistance, microbial genomes, metagenomic data, evolutionary analysis, computational predictions

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Vassiliki Karapapa,
Municipality of Agrinio, Greece

****Correspondence:***
Qi Zhao
zhaoqi@lnu.edu.cn

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

Received: *22 February 2022*
Accepted: *17 March 2022*
Published: *04 April 2022*

Citation:
Wang L, Tay ACY, Li J and Zhao Q (2022) Editorial: Computational Predictions, Dynamic Tracking, and Evolutionary Analysis of Antibiotic Resistance Through the Mining of Microbial Genomes and Metagenomic Data. Front. Microbiol. 13:880967. doi: 10.3389/fmicb.2022.880967

Editorial on the Research Topic

Computational Predictions, Dynamic Tracking, and Evolutionary Analysis of Antibiotic Resistance Through the Mining of Microbial Genomes and Metagenomic Data

Due to the continuous misuse of antibiotics globally, antibiotic resistant arises from antibiotic resistance genes (ARGs) that are now widely detectable from a variety of environmental water and soil resources and various microbial species such as *Escherichia coli* and *Klebsiella pneumoniae* (Von Wintersdorff et al., 2016). As a result, fast and efficient molecular tools are very important to aid the identification, determination, and profiling of antibiotic resistance in environmental samples. In addition, bioinformatics analysis of the microbial genomes and metagenomic data would greatly facilitate our understanding of the molecular mechanisms, environmental transmissions, and dynamic changes of antibiotic resistance (De Abreu et al., 2021). Recently, many advanced bioinformatic methods, including the use of metagenomic next-generation sequencing (Berglund et al., 2019; De Abreu et al., 2021), machine learning (Liu et al., 2020; Anahtar et al., 2021), and Raman spectroscopy (RS) (Tang et al., 2021; Liu et al., 2022), have been proposed to predict ARGs and their mode of action. However, with steady accumulation of massively sequenced data and continuous antibiotic resistance emergence, novel and effective methodologies and tools for ARG prediction and antibiotic resistance profiling analysis and visualization are constantly needed. In order to gain advantage in winning the antibiotic resistance battle, efficient and accurate computational tools are required to determine novel ARGs (Maryam et al., 2021). Under this Research Topic, we sought to highlight an exciting set of groundbreaking efforts proposed by frontline investigators, which mainly focused on implementing computational methodologies to get an in-depth understanding of microbial antibiotic resistance. Articles can be fitted into either of the four categories: (i) novel computational methods, (ii) development of computational tools, (iii) metagenomic data mining, and (iv) microbial genomic analysis. It is expected for some submissions to overlap between categories due to the comprehensive nature of the Research Topic.

In this special issue, majority of the submitted articles were focusing on the novel methods for the rapid and accurate analyses of antibiotic resistance. For example, Wei et al. compared the methods for selecting operational taxonomic units from 16s amplicon sequences. Such research could help biological researchers best select the reasonable clustering method for metagenomic analysis, and facilitate algorithm developers to design more efficient sequence clustering methods. For the discovery and identification of ARGs from fragmented metagenomic assemblies, Shafranskaya et al. presented a novel computational pipeline, termed GraphAMR, to improve read mapping technology. Moreover, Ivanova et al. established a novel bioinformatic pipeline to assist the High-throughput Chromosome Conformation Capture metagenomic analysis, including the identification of bacterial ARGs (or resistomes). Finally, there was a few studies that explored the antibiotic resistance issues from a non-genome-centric angle. For example, Wang et al. summarized recent applications of Raman spectroscopy technique in the antibiotic resistance profiling. They indicated that although there is still a gap between laboratory research and clinical applications for RS, rapid and reliable automatic measurement of the Raman spectra for antibiotic resistance profiling is promising, and eagerly and urgently in need. In another example, Ma et al. developed the Inductive Logistic Matrix Factorization, a novel drug-metabolite association prediction tool that can combine multiple-source interactions between drugs and metabolites and improve prediction performance of drug-metabolite associations, leading to potential applications in the development of novel antibiotics.

Apart from the novel computational methods, two powerful computational pipelines for general analysis of genomes and metagenomes were also presented in this special issue. In brief, Hierarchical Clustering with Kraken (HCK) and Abundance-Base Alternative Approach (ABAA), both developed by Mlaga et al., were designed to classify TS1 amplicons and to detect and filter non-specific amplicons in fungi metabarcoding sequencing datasets, respectively. These two novel pipelines, named HCK-ABAA, had improved the fungi community structures identification and stabilized methodology for metabarcoding analysis. In addition, Hua et al. developed a new web-based server to aid in annotation of AGRs, integrons, and transposable elements. This server could significantly accelerate the bioinformatics analysis of ARG-related sequences.

Two research papers were included in the Research Topic to directly investigate the computational analysis of metagenomic data from clinical perspectives. It is well-known that early, fast, and precise detection of antibiotic resistance is the key to an infection therapy. However, the determination of minimal inhibitory concentrations (MICs) in clinical settings via the conventional agar culturing methods can be very time consuming. To attack this problem, Tan et al., based on the analysis of metagenomic data via XGBoost algorithm and deep neural network (DNN) algorithm, combined single-nucleotide polymorphism (SNP) information and nucleotide k-mers count, and predicted MICs of meropenem against *Klebsiella pneumoniae*. This study significantly improved the ARG detection efficiency. In another study, Han et al. predicted several functional pathways via the computational analysis of fecal microflora composition of acute myocardial infarction (AMI) patients. This could enhance the comprehension of AMI pathogenesis.

Interestingly, a few articles had focused on the geographic distributions and identifications of multi-drug-resistant strains via computational analysis of bacterial genomes. For example, Chung et al. developed a mode-based web tool via Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry to identify multi-drug resistant *Staphylococcus aureus*. In addition, Jiang et al. developed a SNP profiling technique based on whole genome sequencing to facilitate genomic population analyses of *Helicobacter pylori*. This approach holds the potential in understanding the global dissemination of antibiotic resistance genes.

In summary, the findings of the studies collected in this special issue could greatly help mankind in fighting antibiotic resistant in microbial pathogens, from a long-term perspective, and strengthen the faith in finally winning the invisible war all over the world. In addition, we want to thank all the authors who contribute their original work to our special issue and the reviewers for their valuable comments. We would like to express our sincere gratitude to the Specialty Chief Editor, Dr. Matthias Hess and Dr. George Tsiamis, and also the editorial office of Frontier in Microbiology, for their excellent support and providing us with this opportunity to hold this hot topic issue successfully.

AUTHOR CONTRIBUTIONS

LW drafted the manuscript. QZ revised the draft. AT and JL made substantial contributions to the work through in-depth discussion. All authors proposed the Research Topic theme, made a direct and intellectual contribution to the work, and approved the final version for publication.

FUNDING

This study was supported by Foundation of Education Department of Liaoning Province (Grant No. LJKZ0280). LW is a bioinformatician and microbiologist who leads the Theoretical and Experimental Microbiology (TEM) Group https://tem-group.net/ at Xuzhou Medical University.

REFERENCES

Anahtar, M. N., Yang, J. H., Kanjilal, S., and Mcadam, A. J. (2021). Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J. Clin. Microbiol.* 59:e0126020. doi: 10.1128/JCM.01260-20

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. G. J., and Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome* 7:52. doi: 10.1186/s40168-019-0670-1

De Abreu, V. C., Perdigão, J., and Almeida, S. (2021). Metagenomic approaches to analyze antimicrobial resistance: an overview. *Front. Genet.* 11:575592. doi: 10.3389/fgene.2020.575592

Liu, W., Tang, J.-W., Lyu, J.-W., Wang, J.-J., Pan, Y.-C., Shi, X.-Y., et al. (2022). Discrimination between carbapenem-resistant and carbapenem-sensitive *Klebsiella pneumoniae* strains through computational analysis of surface-enhanced raman spectra: a pilot study. *Microbiol. Spectrum* 10:e02409–21. doi: 10.1128/spectrum.02409-21

Liu, Z., Deng, D., Lu, H., Sun, J., Lv, L., Li, S., et al. (2020). Evaluation of machine learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences. *Front. Microbiol.* 11:48. doi: 10.3389/fmicb.2020.00048

Maryam, L., Usmani, S. S., and Raghava, G. P. S. (2021). Computational resources in the management of antibiotic resistance: speeding up drug discovery. *Drug Discov. Today* 26, 2138–2151. doi: 10.1016/j.drudis.2021.04.016

Tang, J.-W., Liu, Q.-H., Yin, X.-C., Pan, Y.-C., Wen, P.-B., Liu, X., et al. (2021). Comparative analysis of machine learning algorithms on surface enhanced raman spectra of clinical Staphylococcus species. *Front. Microbiol.* 12:696921. doi: 10.3389/fmicb.2021.696921

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., et al. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* 7:173. doi: 10.3389/fmicb.2016.00173

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

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

frontiers in Microbiology

METHODS
published: 24 March 2021
doi: 10.3389/fmicb.2021.644012

Comparison of Methods for Picking the Operational Taxonomic Units From Amplicon Sequences

*Ze-Gang Wei[1,2], Xiao-Dan Zhang[1], Ming Cao[3,4], Fei Liu[1], Yu Qian[1] and Shao-Wu Zhang[2]**

[1] Institute of Physics and Optoelectronics Technology, Baoji University of Arts and Sciences, Baoji, China, [2] Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, China, [3] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, [4] School of Mathematics and Statistics, Shaanxi Xueqian Normal University, Xi'an, China

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Wen Zhang,
Huazhong Agricultural University, China
Craig Lee Moyer,
Western Washington University, United States

****Correspondence:***
Shao-Wu Zhang
zhangsw@nwpu.edu.cn

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

Received: *19 December 2020*
Accepted: *17 February 2021*
Published: *24 March 2021*

Citation:
Wei Z-G, Zhang X-D, Cao M, Liu F, Qian Y and Zhang S-W (2021) Comparison of Methods for Picking the Operational Taxonomic Units From Amplicon Sequences. Front. Microbiol. 12:644012. doi: 10.3389/fmicb.2021.644012

With the advent of next-generation sequencing technology, it has become convenient and cost efficient to thoroughly characterize the microbial diversity and taxonomic composition in various environmental samples. Millions of sequencing data can be generated, and how to utilize this enormous sequence resource has become a critical concern for microbial ecologists. One particular challenge is the OTUs (operational taxonomic units) picking in 16S rRNA sequence analysis. Lucky, this challenge can be directly addressed by sequence clustering that attempts to group similar sequences. Therefore, numerous clustering methods have been proposed to help to cluster 16S rRNA sequences into OTUs. However, each method has its clustering mechanism, and different methods produce diverse outputs. Even a slight parameter change for the same method can also generate distinct results, and how to choose an appropriate method has become a challenge for inexperienced users. A lot of time and resources can be wasted in selecting clustering tools and analyzing the clustering results. In this study, we introduced the recent advance of clustering methods for OTUs picking, which mainly focus on three aspects: (i) the principles of existing clustering algorithms, (ii) benchmark dataset construction for OTU picking and evaluation metrics, and (iii) the performance of different methods with various distance thresholds on benchmark datasets. This paper aims to assist biological researchers to select the reasonable clustering methods for analyzing their collected sequences and help algorithm developers to design more efficient sequences clustering methods.

Keywords: operational taxonomic units, 16S rRNA, metagenomics, sequence clustering, high-throughput sequencing

INTRODUCTION

Bacteria constitute an overwhelming majority of domain in the life tree on our planet, occurring in every habitat on earth from natural environments (e.g., oceans, soils, and lakes) to the human body (Sanli et al., 2015; Fuks et al., 2018; Gentile and Weir, 2018). They perform critical functions that range from the regulation of various biogeochemical activities to that of our health and

Abbreviations: AMI, adjusted mutual information; ARI, adjusted rand index (ARI); AL, average linkage; CL, complete linkage; GIS, greedy incremental strategy; MCC, Matthews correlation coefficient; OTUs, operational taxonomic units; rRNA, ribosomal RNA; SL, single linkage; SD,standard deviation.

disease (Shah et al., 2018; Thaiss, 2018; Almeida et al., 2019; Qu et al., 2019a). Describing the taxonomic structure of the communities is vital for studying the bacterial composition and diversity in an environmental or clinical sample (Wei et al., 2016; Lapierre et al., 2019; Zhu et al., 2019). Until recently, most of the bacteria were studied with traditional culture-dependent methods. Because only a small fraction (less than 1%) of all microbial organisms can be isolated, cultivated, and identified in the laboratory, culture-dependent microbial methods are inadequate for exploring the hidden world of many microbial communities (Kellenberger, 2001). On the contrary, metagenomics study is a rapidly growing field that aims to understand all organisms via their nucleic acid sequences to characterize the composition, structure, diversity, and function of microbial communities in a specific habitat (Jo, 2004; Riesenfeld et al., 2004; Laudadio et al., 2019; Wemheuer et al., 2020). Bypassing the needs for isolation and lab cultivation of individual species in traditional microbial studies (Streit and Schmitz, 2004; Meyer et al., 2019), metagenomics allows microbiologists to study the entire genetic materials taken directly from relevant environments and provides a new opportunity to probe the microbial community composition and structure (Koslicki et al., 2013; Zhang et al., 2013; Gao, 2018; Wei and Zhang, 2018; Chong et al., 2020; Qian et al., 2020). Thus, several large-scale metagenomics projects, such as the Human Microbiome Project (Turnbaugh et al., 2007; Integrative HMP (iHMP) Research Network Consortium, 2014), the International Census of Marine Microbes[1], and the Earth Microbiome Project (Gilbert et al., 2014), have been proposed.

In metagenomics, the 16S rRNA (ribosomal RNA) exists in most bacterial species and contains hypervariable regions that allow them to be used as species-specific signatures for identifying taxa (Ward et al., 1990; Stackebrandt and Goebel, 1994; Peterson et al., 2019). Therefore, the 16S rRNA is an ideal proxy for profiling of complex microbial communities and inferring the phylogenetic and evolutionary relations among organisms (Woloszynek et al., 2019). Recently, the rapid advancements in next-generation sequencing (NGS) technologies have dramatically promoted metagenomics studies by offering low-cost and ultra-high-throughput sequencing (Wu et al., 2011). This enormous progress in NGS has resulted in an explosive accumulation of 16S rRNA sequence data (Zhu et al., 2019). How to deal with this massive quantities and high complexity of sequencing data has become a tremendous challenge for microbial researchers (Li et al., 2012; Kim et al., 2013; Qian et al., 2019). As a result, it is needed to develop efficient and accurate computational methods for analyzing these enormous sequence data generated from different habitats and health conditions (Huang et al., 2010; Liu et al., 2014).

Generally, analysis of the 16S rRNA sequencing data typically begins by grouping them into operational taxonomic units (OTUs) (Turnbaugh et al., 2007; Peterson et al., 2009; Větrovský et al., 2018) that contain similar 16S rRNA sequences with high sequence similarity (Seguritan and Rohwer, 2001; Enright et al., 2002; Yooseph et al., 2007; Niu et al., 2010; Westcott and Schloss, 2017). OTUs can represent the microbial taxa and facilitate the downstream analysis for the calculation and visualization of diversity and composition of the microbes (Niu et al., 2011; Zorita et al., 2015; Zou et al., 2018). Thus, picking OTUs has become the backbone in the established workflows, such as QIIME2 (Caporaso et al., 2010; Bolyen et al., 2019), mothur (Schloss et al., 2009), and RDP tools (Wang et al., 2007; Cole et al., 2009, 2013), which are used to analyze the microbial community structures.

In the last decade, a growing number of clustering methods have been proposed to cluster the 16S rRNA sequences into OTUs. However, different methods produce quite diverse outputs, even though a slight parameter change for the same method can also generate distinct results. A more general problem faced by microbial researchers is how to select one suitable method to obtain better clustering results. Therefore, understanding the principle and performance of different clustering algorithms is crucial for users to employ one suitable method for analyzing their sequence data. In this review, we summarized existing state-of-the-art clustering algorithms, explained their clustering mechanisms, analyzed their characters, compared their clustering performance on several benchmark datasets, and recommended some directions for developing new clustering algorithms. We hope this review can assist the biological researchers to select a reasonable clustering method for analyzing their collected sequences and help algorithm developers to design more efficient sequence clustering methods.

METHODS OF OPERATIONAL TAXONOMIC UNIT PICKING

Numerous OTU picking methods have been developed, which can be categorized as closed-reference clustering, *de novo* clustering (also called taxonomy independent), and open-reference clustering (Lawley and Tannock, 2017; Whelan and Surette, 2017; De Filippis et al., 2018). The closed-reference clustering involves comparing each query sequence to an annotated reference taxonomy database by utilizing the sequence classification or searching methods (Liu et al., 2017, 2018; Matias Rodrigues et al., 2017; Wei et al., 2020), then sequences matched to the same reference sequence are grouped into the same OTU. However, if a large portion of microbes in a sample has not yet been well defined, that is, not recorded in databases (i.e., unknown taxa), then they cannot be assigned to an OTU. Thus, closed-reference clustering methods are largely dependent on the completeness of the reference database, hence, have a poor performance on the condition that many novel organisms exist in the sequencing data (Schloss and Westcott, 2011; Chen et al., 2016). Furthermore, two query sequences matched to the same reference sequence may have a lower similarity to each other (Westcott and Schloss, 2015). As a result, closed-reference methods are often applied for the purpose of sequence annotation (Sun et al., 2011). For *de novo* clustering, all sequences are clustered into OTUs based on the pairwise sequence distances rather than comparing against a reference database (Forster et al., 2016). That is, *de novo* clustering methods compare each sequence against each other, followed by implementing different

[1] http://icomm.mbl.edu

clustering algorithms at a specified threshold to group sequences into OTUs. For the open-reference clustering, it is a combination of the closed-reference and *de novo* methods. Here, a closed-reference clustering approach is first used to assign OTUs, and the unassigned sequences outputted by the closed-reference approach are then grouped by a *de novo* clustering method. Open-reference clustering blends the strengths and weaknesses of the other method and adds the complication that closed-reference and *de novo* clustering use different OTU definitions (Westcott and Schloss, 2017). As a result, *de novo* clustering does not depend on any reference database and, hence, can assign all sequences into OTUs, including both sequences that have been deposited in annotated databases as well as novel unknown ones (Zou et al., 2018). Additionally, several studies (Jackson et al., 2016; Schloss, 2016) also show that *de novo* clustering methods significantly outperform the other two approaches for picking OTUs. Therefore, *de novo* clustering attracts more attention and has become the preferred choice for researchers (Schloss, 2010; Cai et al., 2017). In the following, we mainly focus on *de novo* clustering.

Many different *de novo* clustering methods have been proposed to pick OTUs in the past decade, which can be further classified into four general categories: hierarchical clustering, heuristic clustering, model-based, and network-based clustering methods.

Hierarchical Clustering Methods

Hierarchical clustering methods generally require a full distance matrix between all sequences based on pairwise sequence alignment or multiple sequence alignment, then construct a hierarchical tree on the distance matrix. By applying a predefined clustering threshold to the hierarchical tree, sequences within the threshold are grouped into one OTU, as shown in **Figure 1**. Actually, most hierarchical methods implement the complete-linkage (CL), average-linkage (AL), or single-linkage (SL) algorithms (Zhang and Wei, 2015). CL, SL, and AL belong to the agglomerative methods, that is, in the beginning, each sequence is one cluster, then compute the similarity (i.e., distance) between each of the clusters and merge the two most similar clusters. Repeat the previous step until there is only a single cluster left, or the merging distance meets the given threshold (**Figure 1C**). The main differences among CL, SL, and AL are the distance criteria defined between two clusters (**Figure 2**), which can reflect the degree of clustering. For SL, the distance between two clusters is the minimum distance between two sequences in each cluster (**Figure 2A**). For CL, the distance between two clusters is defined as the maximum distance between two sequences in each cluster (**Figure 2B**). For AL, the distance between two clusters is defined as the average distance between each sequence in one cluster to every sequence in the other cluster (**Figure 2C**). We can see that SL is a loose clustering strategy, CL is the most stringent, and AL is the middle ground between SL and CL.

DOTUR (Schloss and Handelsman, 2005) is probably the first published tool for hierarchically clustering sequences into OTUs by using CL, AL, and SL. mothur (Schloss et al., 2009), the improved version of DOTUR, has become the representative hierarchical clustering method for picking OTUs. As with DOTUR, mothur needs to load the distance matrix into computer memory before performing clustering. In order to alleviate the computational complexity and memory usage, Sun et al. (2009) proposed a novel algorithm (namely, ESPRIT), which adopts the *k*-mer (substrings of length *k*) distance to rapidly identify the sequence pairs with high similarity and stores the reduced distance by using a sparse matrix. In the procedure of picking OTUs, the Hcluster algorithm was devised to perform CL clustering, which can avoid loading the whole matrix into memory. Huse et al. (2010) observed that the CL algorithm is sensitive to sequencing artifacts, then they proposed a single-linkage preclustering (SLP) to overcome the effect of sequencing errors and decrease the inflation of OTUs. Cole et al. (2013) proposed the mcClust algorithm to achieve the CL strategy that allows the distance matrix computation to be parallelized, which can lower the time complexity. Matias Rodrigues and von Mering (2013) presented the HPC-CLUST pipeline, a distributed implementation of two hierarchical clustering algorithms (CL and AL) with high optimization. HPC-CLUST takes as input a set of pre-aligned sequences and efficiently allocates both memory usage and computing complexity, which can handle large numbers of sequences on a computer cluster. Franzén et al. (2015) developed the oclust method in which the distance matrix and CL clustering are performed with an R package based on the pre-aligned sequences. Similar to the HPC-CLUST, the oclust also needs to pre-align sequences, which is usually computation intensive.

Generally, the computational complexity of hierarchical algorithms both in time and space is $O(N^2)$, where N is the number of sequences. Thus, the computational cost of most hierarchical methods quadratically scales with the number of sequence increases. As a result, hierarchical clustering methods are not suitable for handling huge numbers of sequences because of their intrinsic computing complexity (Barriuso et al., 2011).

Heuristic Clustering Methods

Heuristic clustering processes input sequences one by one, avoiding the expensive step of computing distances of all pairwise sequences. Most classical heuristic clustering methods use pairwise sequence alignment and generate clusters in a greedy incremental strategy (GIS), which is shown in **Figure 3**. These methods use one sequence (called seed) to represent its cluster, and each query sequence is compared with all seeds of existing clusters (Chen et al., 2016). One query sequence is assigned to a cluster if the distance between the sequence and one seed meets the clustering threshold (**Figure 3A**). Otherwise, a new cluster is created, and the query sequence becomes the seed sequence (**Figure 3B**). Due to the comparison of all sequences just with the seeds of clusters, greedy heuristic clustering is computationally much more efficient than hierarchical clustering methods. As a result, many different heuristic clustering algorithms have been developed, and the main differences are the seed selection and distance calculation.

CD-HIT (Li and Godzik, 2006; Huang et al., 2010; Fu et al., 2012) and USEARCH (Edgar, 2010) are the two best-known heuristic methods for picking OTUs. The main discrepancy between these two methods is the sequence sorting before

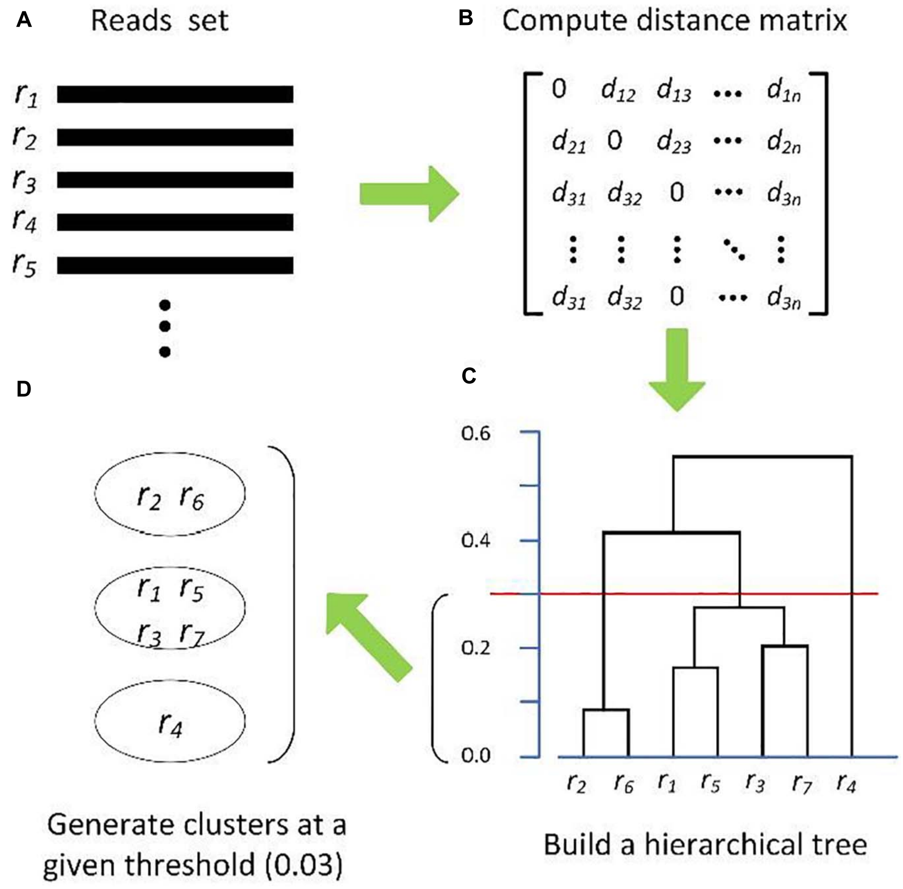

A
Reads set
r_1
r_2
r_3
r_4
r_5
B
Compute distance matrix
0 d_{12} d_{13} ••• d_{1n}
d_{21} 0 d_{23} ••• d_{2n}
d_{31} d_{32} 0 ••• d_{3n}
d_{31} d_{32} 0 ••• d_{3n}
C
0.6
0.4
0.2
0.0
r_2 r_6 r_1 r_5 r_3 r_7 r_4
Build a hierarchical tree
D
r_2 r_6
r_1 r_5 r_3 r_7
r_4
Generate clusters at a given threshold (0.03)

FIGURE 1 | Schematic diagram of hierarchical clustering algorithms. **(A)** Input reads set, **(B)** distance matrix, **(C)** hierarchical Tree, and **(D)** OTUs formation.

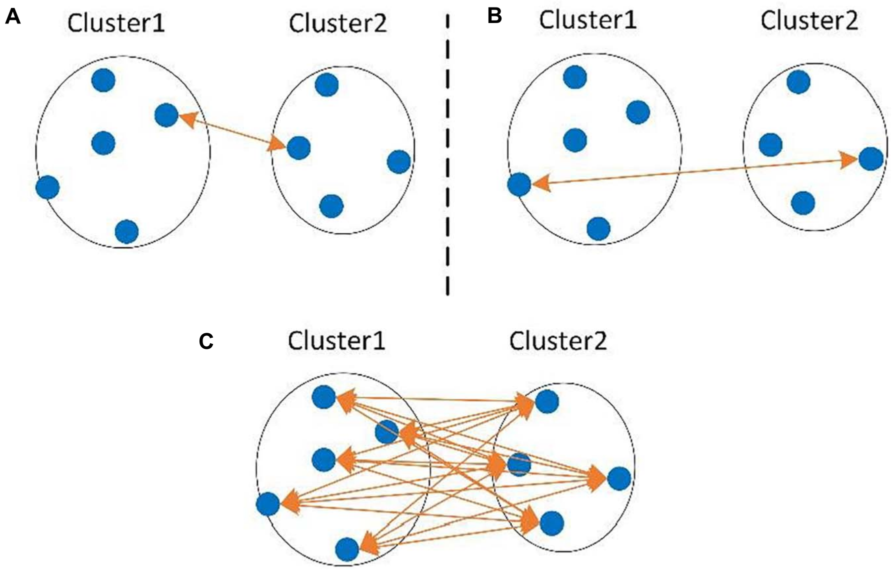


FIGURE 2 | The distance between two clusters defined in single-linkage (SL) **(A)**, complete-linkage (CL) **(B)**, and average-linkage (AL) **(C)** clustering algorithms.

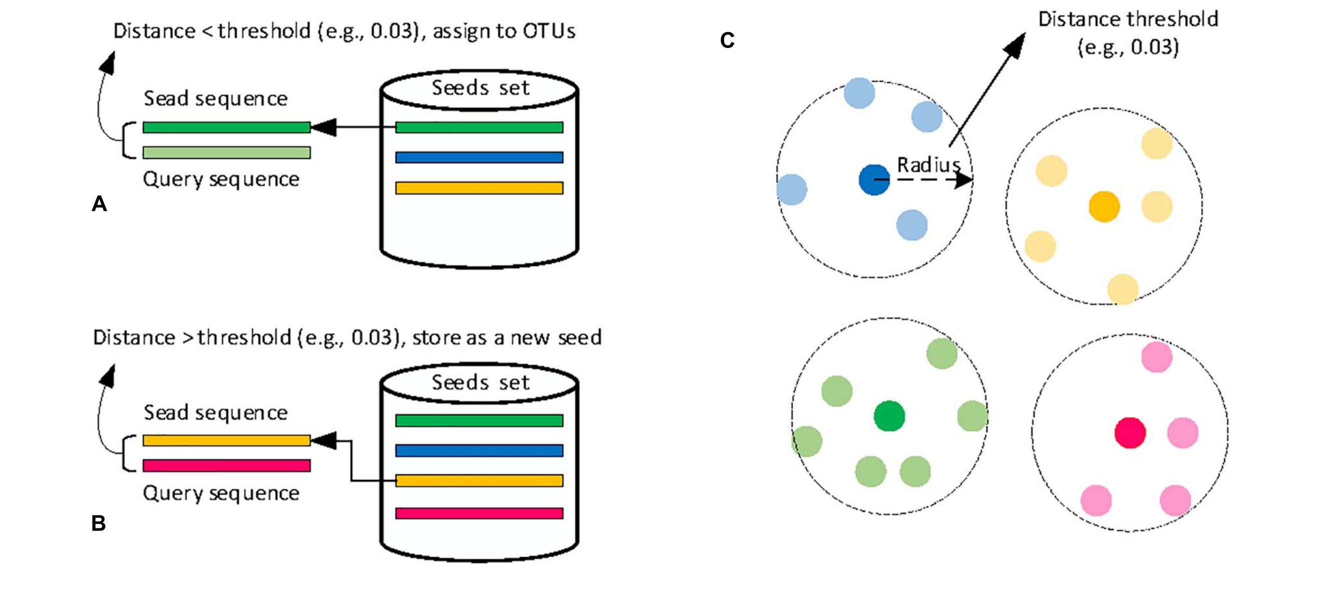


FIGURE 3 | Schematic diagram of classical heuristic clustering methods. **(A)** sequence assignment, **(B)** new seed generation, and **(C)** OTUs results.

clustering. CD-HIT sorts by the length of sequences while USEARCH by sequence abundance. UPARSE (Edgar, 2013) is an improved version of USEARCH, which adds the chimera detection for seed sequences. Different from sequence distance calculation in CD-HIT and USEARCH, GramClust (Russell et al., 2010) designs a distance metric based on the inherent grammar of each pairwise sequences for clustering a set of sequences. DNACLUST (Ghodsi et al., 2011) also follows the GIS way, but it uses a novel *k*-mer-based filtering algorithm to accelerate the clustering procedure. Similar to DNACLUST, LST-HIT (Namiki et al., 2013) introduces a new filtering scheme to remove dissimilar sequence pairs on the basis of the longest common subsequence before performing pairwise sequence alignment, which can speed up the computation. SUMACLUST (Mercier et al., 2013) and OTUCLUST (Albanese et al., 2015) are another two greedy clustering methods that are designed to perform exact sequence alignment, rather than semiglobal alignments implemented in CD-HIT and USEARCH. Additionally, OTUCLUST performs sequence de-duplication and chimera removal. LSH (Rasheed et al., 2013) is also another greedy clustering algorithm that utilizes the locality-sensitive hashing to accelerate the pairwise sequence comparisons and incorporates a matching criterion to improve the quality of sequence comparisons. Considering that using a single global clustering threshold is too relaxed for slow-evolving lineages, Mahé et al. (2014) designed Swarm, which first generates an initial set of OTUs by iteratively agglomerating similar sequences, then breaks them into sub-OTUs to refine the clustering results by using abundance information and OTUs' internal structures. VSEARSH (Rognes et al., 2016) is a free 64-bit and open-source versatile program and is designed as an alternative to the USEARCH tool for which the source code is not publicly available and only a memory-confined 32-bit version is freely available for academic users.

The above heuristic methods just select one sequence as the seed to represent the cluster. Once the seed is selected, it will not be changed anymore, resulting in the outcomes sensitive to the selected seeds. Therefore, how to select a "good" seed that includes more cluster information is significantly important. Some methods have been proposed to achieve this target. Zheng et al. (2012) introduced a dynamic seed-based clustering method (namely, DySC) to reselect seed sequences. DySC first uses the traditional GIS to form the pending clusters. Once a pending cluster reaches a threshold size, it is converted into a fixed cluster, and a new fixed seed is reselected, which is defined as the sequence that maximizes the sum of *k*-mers shared between the fixed read and other reads in one cluster. Chen et al. (2013a) proposed MSClust, a multiseed-based heuristic clustering method. The multiseeds for one cluster are generated based on an adaptive strategy, that is, one query sequence is assigned to one cluster if the average distance between the sequence and seeds is smaller than the user-defined threshold; otherwise, the sequence is marked as unassigned. In order to reduce the sensitivity of seeds to sequencing errors, we developed DBH (Wei and Zhang, 2017), a de Bruijn (DB) graph-based heuristic clustering method. It first forms temporary clusters using the traditional GIS. When the size of a temporary cluster reaches the predefined minimum sequence number, DBH builds a DB graph for this cluster and generates a new seed to represent this cluster. Finally, the remaining sequences are assigned to the corresponding OTUs. Later, We designed DMSC (Wei and Zhang, 2019), a dynamic multiseed clustering method for OTU picking. DMSC first generates a series of clusters based on the GIS strategy. When the sequence number in a cluster is larger than the value of a predefined size, the multicore sequence (MCS) selection procedure is triggered, and the MCS is applied as the seeds of the cluster. The MCS is determined as the *n*-core sequences ($n \geq 3$) that the distance between any two sequences in the MCS is less than the clustering threshold. If a new sequence is added to one cluster according to the average distance to MCS and the distance standard deviation in MCS, DMSC will update the MCS. By reselecting seed sequences, these four methods can achieve higher clustering accuracy than the traditional heuristic methods such as CD-HIT and USEARCH. Recently, Bazin et al. (2018) proposed a fuzzy OTU-picking algorithm that adds the uncertainty information to the clustering based on fuzzy sets, which can also improve the clustering quality.

Different from most existing clustering methods that use the seed sequences to represent clusters, Cai and Sun (2011) developed the ESPRIT-Tree method, which initially constructs a PBP (pseudometric-based partition) tree that provides a coarse representation of the entire sequences, then iteratively finds the closest pairs of sequences or clusters and merges them into a new cluster. Later, they proposed an improved method of ESPRIT-Forest (Cai et al., 2017), which can cluster massive sequence data in a subquadratic computational complexity. Pagni et al. (2013) introduced DBC454 for clustering ITS1 (fungal internal transcribed spacer 1) sequences using a density-based hierarchical clustering procedure. Recently, Westcott and Schloss (2017) designed OptiClust that maximizes the value of Matthews correlation coefficient (MCC) by iteratively reassigning sequences to new OTUs.

Broadly speaking, heuristic clustering methods have a lower computational complexity of $O(KN)$, where K is the final number of clusters. Usually $K \leq N$, and hence, heuristic clustering methods are computationally much more efficient than hierarchical clustering methods and are more widely employed to deal with hundreds of thousands of 16S rRNA sequences.

Model-Based Clustering Methods

One of the critical problems with most existing hierarchical and heuristic clustering methods is the need to select a constant and optimal distance threshold to define OTUs at a distinct taxonomic level (e.g., species). A slight change in threshold can result in very different OTUs. Model-based clustering methods, such as CROP (Hao et al., 2011), BEBaC (Cheng et al., 2012), and BC (Jääskinen et al., 2014), were developed to address this issue. CROP (Hao et al., 2011) builds a Bayesian model to cluster sequences, which utilizes a Gaussian mixture model and a birth–death process to characterize a specific cluster. BEBaC (Cheng et al., 2012) first uses the heuristic trick to assign the highly similar sequences to form a pregroup, then similar 3-mer count vectors are assigned into crude clusters by searching for the best partitions that achieve the maximum posterior possibility for given sequence data. In the fine clustering phase, BEBaC applies a minimum description length criterion to determine the number of OTUs, generating the final partitioning. BC (Jääskinen et al., 2014) first models the sequences using Markov chains, then uses a Bayesian partition model with the Dirichlet process to split and merge clusters. Although these methods partition sequences into OTUs without additional information besides the sequence data itself, it is not suitable for large-scale sequence datasets.

Network-Based Clustering Methods

Several network-based clustering methods such as M-pick (Wang et al., 2013), MtHc (Wei and Zhang, 2015), and DMclust (Wei et al., 2017) were also proposed to solve the problem of requiring a given clustering distance to pick OTUs. **Figure 4** shows the schematic diagram of the main processing steps in network-based clustering methods. M-pick (Wang et al., 2013) first compute the distances across all pairs of sequences to construct a fully connected graph, then prunes the complete graph to generate a neighborhood graph; finally, a modularity-based community detection approach is recursively performed to form OTUs. Based on the concept of network motif, we proposed MtHc (Wei and Zhang, 2015). MtHc first searches for sequence motifs using a heuristic strategy then uses these sequence motifs as seeds to generate candidate clusters, which are hierarchically merged into OTUs based on the distances of motifs between two clusters. Later, we developed DMclust (Wei et al., 2017); it first searches for the sequence dense groups, which are viewed as nods to construct a weighted graph, then a modularity-based clustering method is applied to capture the community structures in sequence data to generate clusters.

Network-based methods require a full distance matrix of all pairwise sequences to construct a graph and, hence, has a high computational complexity in terms of run time and memory usage. They cannot handle large numbers of sequences.

Based on the above analysis, **Figure 5** describes the development history of clustering methods according to their published years. It can be summarized that hierarchical clustering (either based on AL, SL, or CL) and network-based clustering methods need to compute and store a full distance matrix of all pairwise sequences, adding the computational complexity and memory space usage. Although the model-based clustering method could produce better clustering results, their run time would render them unusable on massive quantities of sequences. Due to the comparison of each sequence just with the seed sequences, heuristic clustering methods are capable of handling millions of sequences and are more widely employed to analyze massive 16S rRNA datasets (Cai and Sun, 2011). With the sequencing technology development, the volume of sequences increases drastically, and heuristic clustering methods continue to attract more attention in picking OTUs.

MATERIALS OF BENCHMARK DATASETS AND EVALUATION METRICS

Benchmark Datasets

Three benchmark studies, including one simulated and two real-world sequence datasets, were conducted to assess the performance of 12 existing OTU-picking algorithms. The simulated dataset was directly produced by Seq-Gen (Rambaut and Grass, 1997) sequence simulator. It can be directly downloaded from BEBaC (Cheng et al., 2012). Two real-life sequence datasets are the V4 hypervariable region dataset from the murine gut and the global 16S bacterial rRNA gene sequence dataset, respectively. These sequence datasets have been widely

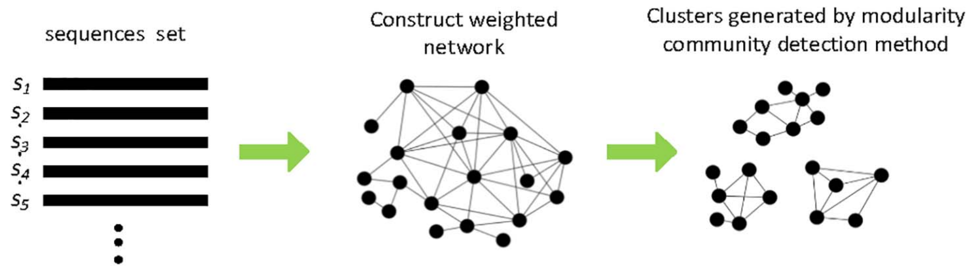

sequences set
s_1
s_2
s_3
s_4
s_5
Construct weighted network
Clusters generated by modularity community detection method

FIGURE 4 | Schematic diagram of network-based methods.

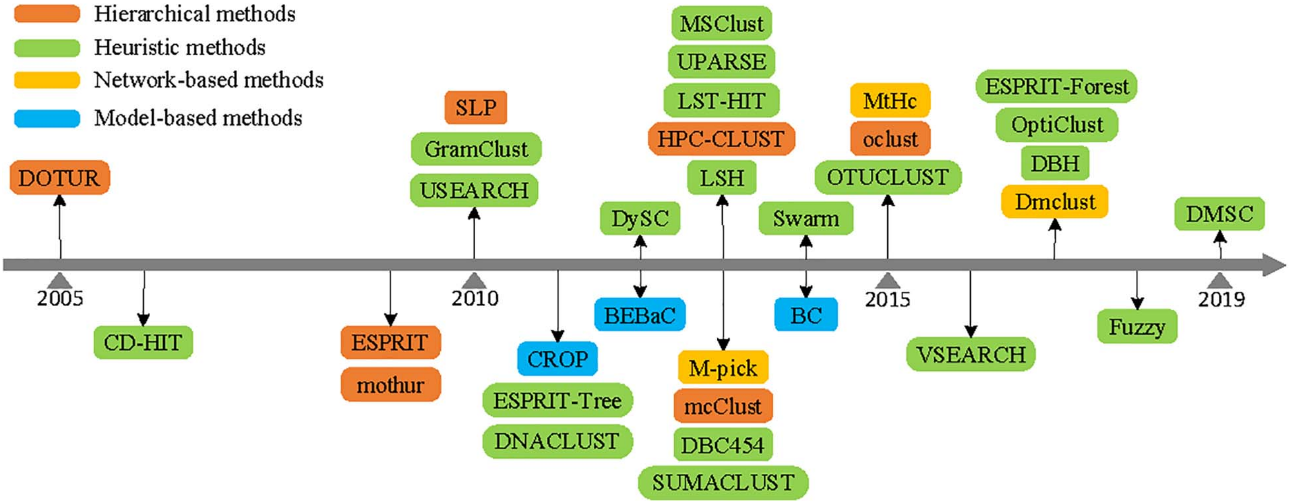


FIGURE 5 | Published years of operational taxonomic unit (OTU) picking methods (mentioned in this paper).

used to validate the performance of clustering results (Cheng et al., 2012; Wei and Zhang, 2017, 2019).

For the simulated dataset, the ground truths (labels of sequences) are directly taken from simulated data, in which we exactly know the species of each sequence. However, for real-life datasets, we need to construct the ground-truth information by searching a reference database. The processing procedures of obtaining ground truth information for real-life datasets are described in **Supplementary Figure S1**. First, the V4 pair-end sequencing data are merged by the FLASH (Magoè and Salzberg, 2011) assembly tool. Then, the merged sequences are cleaned to remove sequences with low quality and short length by quality USEARCH (Edgar, 2010) filtering software. The Python executive command (*assign_taxonomy.py*) in QIIME (Caporaso et al., 2010) is applied to align the cleaned sequences to the default reference database (Greengenes DeSantis et al., 2006) to obtain the species information. Last, aligned sequences with high alignment quality (i.e., >97% identity over an aligned region >90% of the length of the sequences) are retained, and the remaining annotated sequences are adopted to construct the final ground-truth. These procedures of constructing the ground-truth information are based on previous studies (Cai and Sun, 2011; Wei and Zhang, 2019). Some detailed features (such as taxon number, sequences number, and average sequence length) of three benchmark datasets are listed in the following **Table 1**.

Evaluation Metrics

The number of OTUs, normalized mutual information (NMI), Matthews correlation coefficient (MCC), adjusted rand index (ARI), and adjusted mutual information (AMI) metrics are used to evaluate the clustering performance. OTU number is the cluster number that directly inflects the count of species (or genera). NMI value is commonly applied to estimate the clustering accuracy, that is, how the outcome of one clustering algorithm agrees with the ground truth (Chen et al., 2013b). ARI (Nguyen et al., 2015; Jin and Bi, 2018) represents the number of pairwise sequences that are either in the same cluster or in different clusters in both partitions. AMI is similar to ARI. Different from NMI, AMI, and ARI that rely on an external reference, the metric of MCC can be calculated according to the clustering threshold and distances between sequences (Schloss and Westcott, 2011); thus, MCC is regarded as an objective criterion to evaluate the clustering quality of different algorithms for OTU picking (Westcott and Schloss, 2015; Schloss, 2016; Liu et al., 2019). AMI, ARI, and MCC vary

TABLE 1 | Statistics of three benchmark datasets for operational taxonomic unit (OTU) picking.

Sequence data	Taxon number	Total sequences	Average length	Variable regions	References
Simulated dataset	9	22 K	500	-	Cheng et al., 2012
V4 dataset	68	∼511 K	253	V4	Westcott and Schloss, 2015
Global 16S rRNA	1,498	∼887 K	∼1,400	V1-V9	Matias Rodrigues and von Mering, 2013

between -1 and 1, and a larger value represents better clustering quality. How to calculate these metrics are provided in the **Supplementary File**.

COMPARISON RESULTS

We evaluate 12 state-of-the-art OTU picking methods, that is, CD-HIT (v.4.6.8) (Li and Godzik, 2006), USEARCH (v.11.0.667) (Edgar, 2010), DNACLUST (Ghodsi et al., 2011), Swarm (v.1.2.19) (Mahé et al., 2014), VSEARCH (v.2.3.4) (Rognes et al., 2016), DBH (Wei and Zhang, 2017), DMSC (Wei and Zhang, 2019), DySC (v.06-1-2012) (Zheng et al., 2012), ESPRIT-Forest (Cai et al., 2017), GramClust (v.1.3) (Russell et al., 2010), average linkage (AL) clustering method employed in mothur software (v.1.44.3) (Schloss et al., 2009), and CROP (Hao et al., 2011). Among these methods, CD-HIT, USEARCH, DNACLUST, Swarm, VSEARCH, DySC, ESPRIT-Forest, DBH, GramClust, and DMSC are the typical heuristic clustering methods; mothur is a comprehensive software package for sequence clustering, and it is demonstrated that the AL clustering implemented in mothur (mothur-AL) is a reliable method to represent the actual distances between sequences (Westcott and Schloss, 2015); CROP is a model-based method. All methods were executed on the same Linux server for OTU picking. The running parameters and command lines of each algorithm are given in **Supplementary Table S1**.

Benchmarking on the Simulated Dataset

Figure 6 shows the NMI values of 12 clustering methods as a function of distance thresholds ranging from 0.01 to 0.1. Because Swarm does not apply the distance threshold to cluster, and just uses the parameter d (d nucleotide differences) to generate OTUs, the setting of d is calculated by $d = d_{\mathrm{th}} \times L_{\mathrm{ave}}$, where L_{ave} is the average length (i.e., 500) of this simulated data, d_{th} is the distance threshold ranging from 0.01 to 0.1. From **Figure 6**, we can see that all methods, except VSEARCH and GramClust, show a similar trend, that is, they achieved higher NMI values near 0.04 distance but lower NMI when the distance threshold increases. The NMI peak values of the different methods occur at different distance thresholds. This is mainly due to the discrepancies of distance calculation and clustering strategy in each method. VSEARCH shows a different trend from other methods. It obtained the NMI peak at 0.07 distance, while the other methods achieved their NMI peak value near 0.04 distance. The NMI values of GramClust is always between 0.85 and 0.90 even in lower distances. The peak NMI scores of 11 methods and the corresponding inferred OTU number at different distance thresholds are reported in **Table 2**. It can be found that DMSC, CROP, DBH, CD-HIT, VSEARCH, DNACLUST, Swarm, GramClust, and mothur-AL successfully generated nine OTUs at their maximum NMI values, while USEARCH, DySC, and ESPRIT-Forest overestimated OTUs.

Figure 7 illustrates the MCC values of 12 OTU picking methods at different clustering thresholds. Similar to the NMI curve, all methods achieved the highest MCC value near 0.04 distance threshold, while USEARCH and VSEARCH obtained their MCC peak values at 0.01 distance. **Table 3** reports the average, standard deviation (SD), and maximum of MCC scores with the inferred OTUs number. It can be observed that DMSC, CROP, Swarm, GramClust, DBH, and mothur-AL methods also can produce the exact OTU number at their best MCC values, while USEARCH, DySC, ESPRIT-Forest, CD-HIT, VSEARCH, and DNACLUST overestimated the OTU number. Based on the MCC values listed in **Table 3**, we can see that DMSC, ESPRIT-Forest, CD-HIT, and mothur-AL have a better clustering quality (ave. MCC > 0.9) than other methods, and mothur-AL has the best average MCC value. The NMI values, OTUs number, and MCC values of the different methods in the range of 0.01–0.1 distance thresholds can be seen in **Supplementary Table S2**.

Supplementary Figures S2, S3 depict the ARI and AMI curves of 12 OTU picking methods at different clustering thresholds. On the whole, the curves of ARI and AMI are similar to those of NMI. That is, most methods, e.g., CD-HIT, DBH, DySC ESPRIT-Forest, DNACLUST, Swarm, DMSC, and mothur-AL obtained higher ARI and AMI values near 0.04 distance but lower ARI when the distance threshold increases, while VSEARCH and UCLUST show a different trend from other methods where they obtained the ARI peak at 0.07 distance. The ARI values of GramClust are always between 0.65 and 0.67 even in lower distances, and AMI values are between 0.79 and 0.81. Although CROP achieved the highest ARI (at 0.01 distance threshold) among all methods, it generated 158 OTUs, 17 times larger than the true number. The maximum ARI and AMI values of the 11 methods at different clustering thresholds are listed in **Supplementary Tables S3, S4**. It can be found that some clustering methods (such as DMSC, VSEARCH, DNACLUST, Swarm, GramClust, DBH, and mothur-AL) can exactly infer the true number of OTUs at their best ARI and AMI values for the simulated dataset.

Benchmarking on V4 Dataset

For the V4 dataset, just eight methods of USEARCH, CD-HIT, DBH, GramClust, DNACLUST, VSEARCH, DMSC, and mothur-AL can generate the clustering results at each distance threshold, while ESPRIST-Forest, DySC, CROP, and Swarm cannot handle this dataset. **Figure 8** shows the NMI curves of each clustering

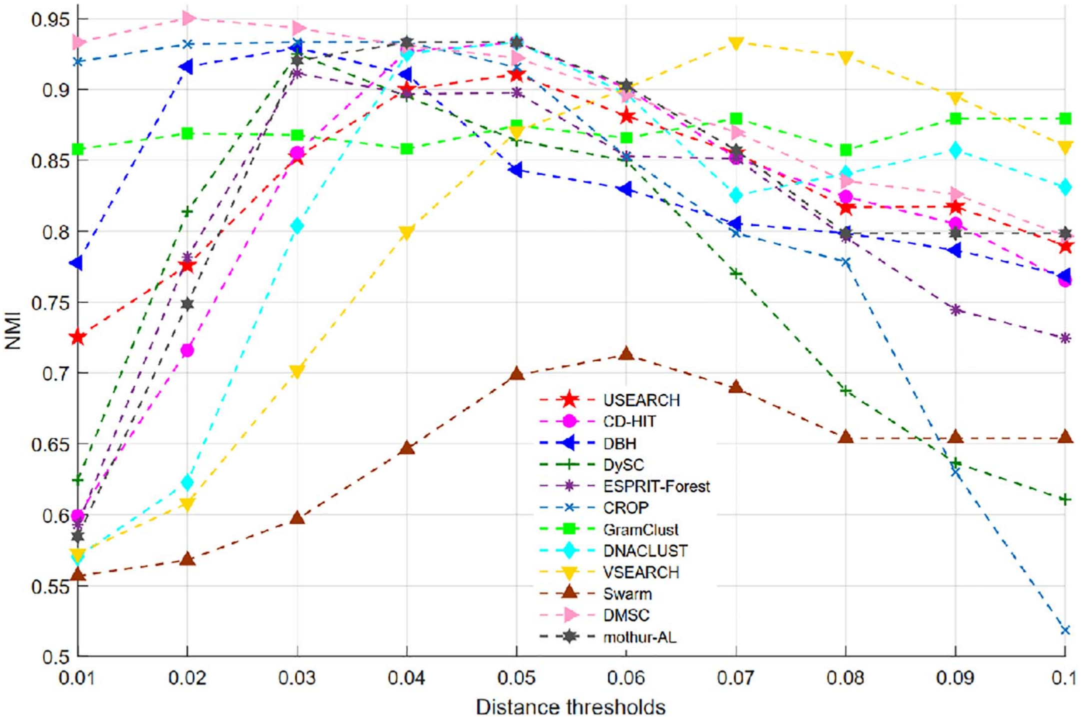


FIGURE 6 | Normalized mutual information (NMI) values of different clustering methods on the simulated dataset.

TABLE 2 | Maximum normalized mutual information (NMI) values for different OTU picking methods on the simulated dataset.

	DMSC (0.02)	USEARCH (0.05)	DySC (0.03)	ESPRIT-Forest (0.05)	CD-HIT (0.05)	CROP (0.03)
Max. NMI	0.9503	0.9107	0.9252	0.8979	0.9334	0.9334
OTUs number	9	10	17	13	9	9
	VSEARCH (0.07)	**DNACLUST (0.05)**	**Swarm (*d* = 15)**	**GramClust (0.07)**	**DBH (0.03)**	**Mothur-AL (0.04)**
Max. NMI	0.9334	0.9333	0.9334	0.8795	0.9293	0.9333
OTUs number	9	9	9	9	9	9

The value in the parentheses is the clustering threshold where each method achieves its peak NMI. For the Swarm method, it is the value of parameter d.

method, and **Supplementary Figure S4** presents the inferred OTU number at different clustering thresholds. We can see that GramClust has higher NMI scores than other approaches when the distance increases from 0.01 to 0.06. DMSC and mothur-AL have higher NMI values than the other methods at distance thresholds from 0.09 and 0.11, and mothur-AL achieved the highest NMI score at 0.12 threshold. For the OTU number in **Supplementary Figure S4**, all methods show a similar descending trend from 0.01 to 0.15, generating close OTU number to the ground truth near 0.1 distance except GramClust and mothur-AL. mothur-AL obtained close OTU number at 0.08 distance threshold. GramClust produces more OTUs than the ground truth even in low distance thresholds. The ARI and AMI curves of each clustering method are described in **Supplementary Figures S5, S6**, which show a similar result to the curve of NMI.

Figure 9 describes the MCC values at different distance thresholds, and **Table 4** reports the maximum, average, and SD of MCC values for each method. Obviously, from **Table 4**, we can find that DMSC, DNACLUST, and mothur-AL achieve higher average MCC values than other clustering methods, indicating that these three methods can produce higher clustering quality on the V4 dataset. The NMI values, OTU number, MCC, ARI, and AMI values of each method with different distance thresholds can be found in **Supplementary Tables S5, S6**.

Benchmarking on Global 16S rRNA Sequence Dataset

The global 16S rRNA dataset was often employed to test the scalability of dealing with longer sequences. For this near full-length 16S dataset, only USEARCH, CD-HIT, VSEARCH, and DBH can get the clustering results. Other methods fail to hand with this large-scale dataset.

The NMI values of USEARCH, CD-HIT, VSEARCH, and DBH with different clustering thresholds are shown in **Supplementary Figure S7**. We can observe that CD-HIT

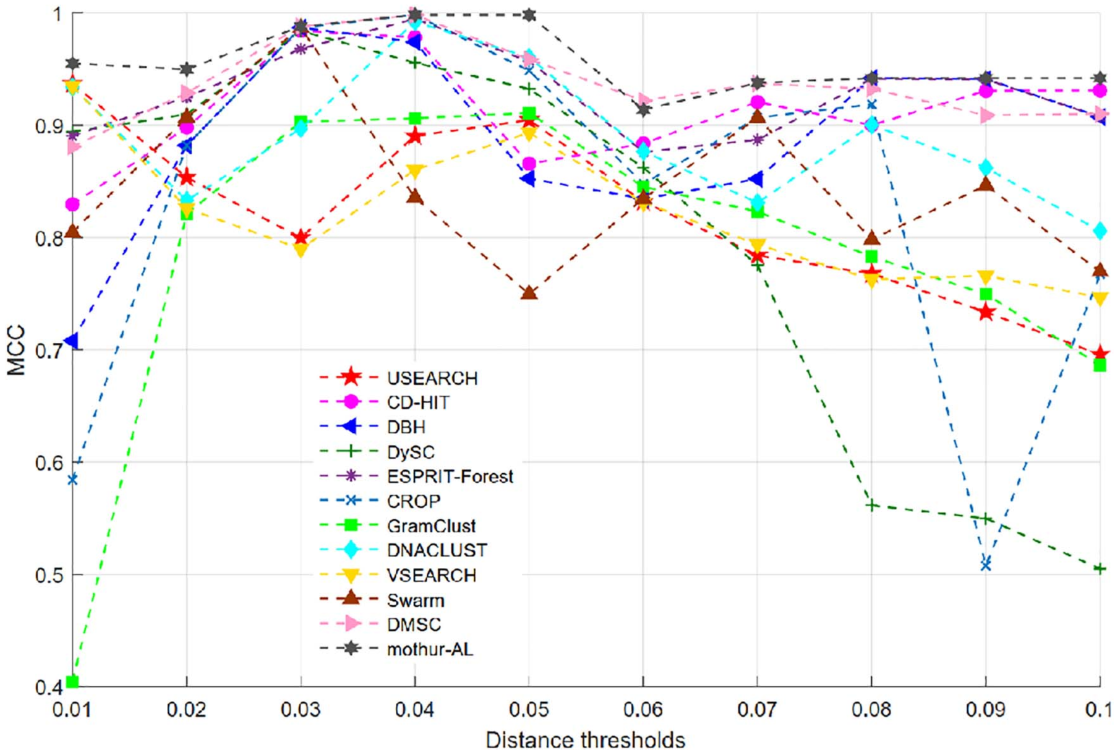


FIGURE 7 | The Matthews correlation coefficient (MCC) values of 12 OTU picking methods on the simulated dataset.

TABLE 3 | The average, SD, and maximum MCC values of 11 OTU picking methods on the simulated dataset.

	DMSC (0.04)	USEARCH (0.01)	DySC (0.03)	ESPRIT-Forest (0.04)	CD-HIT (0.03)	CROP (0.03)
Max. MCC	0.9980	0.9369	0.9838	0.9947	0.9840	0.9980
OTUs number	9	528	17	16	27	9
Ave. MCC	0.9363	0.8198	0.7929	0.9286	0.9120	0.8347
SD of MCC	0.0343	0.0737	0.1750	0.0366	0.0451	0.1585
	VSEARCH (0.01)	**DNACLUST (0.04)**	**Swarm (*d* = 15)**	**GramClust (0.05)**	**DBH (0.03)**	**Mothur-AL (0.04)**
Max. MCC	0.9349	0.9921	0.9868	0.9106	0.9868	0.9980
OTUs number	1,291	15	9	9	9	9
Ave. MCC	0.8204	0.8891	0.5474	0.7832	0.8879	0.9564
SD of MCC	0.0578	0.0567	0.1385	0.1436	0.0781	0.0270

The value in the parentheses is the clustering threshold where each method achieves its peak MCC. For the Swarm method, it is the value of parameter d.

achieves higher NMI scores than other approaches at distance thresholds from 0.01 to 0.07, while USEARCH and VSEARCH obtain higher NMI scores than DBH and CD-HIT with distance increases from 0.11 to 0.15. The AMI values of USEARCH, CD-HIT, VSEARCH, and DBH are described in **Supplementary Figure S8**, which shows a similar result to the NMI values in **Supplementary Figure S7**. **Supplementary Figure S9** represents the OTU number inferred by these four methods. It can be seen that four OTU picking methods present a similar trend, that is, the OTU number exponentially decreases when the clustering distance increases. Four OTU picking methods of USEARCH, CD-HIT, VSEARCH, and DBH overestimate OTUs in the distance range from 0.01 to 0.13. **Supplementary Figure S10** shows the ARI values of USEARCH, CD-HIT, VSEARCH, and DBH. We can see that CD-HIT achieves higher ARI values than other methods at distance thresholds from 0.01 to 0.07, DBH obtains the highest ARI at distance thresholds from 0.08 to 0.10, while USEARCH and VSEARCH obtain higher NMI scores than DBH and CD-HIT with distance ranging from 0.12 to 0.15. **Supplementary Figure S11** describes the MCC values of four OTU picking methods. Obviously, DBH achieves higher MCC values than CD-HIT, USEARCH, and VSEARCH at any distance threshold, indicating that DBH can produce better clustering quality for this full-length 16S rRNA dataset. The NMI, MCC, ARI, AMI values, and OTU number of each method are provided in **Supplementary Table S7**.

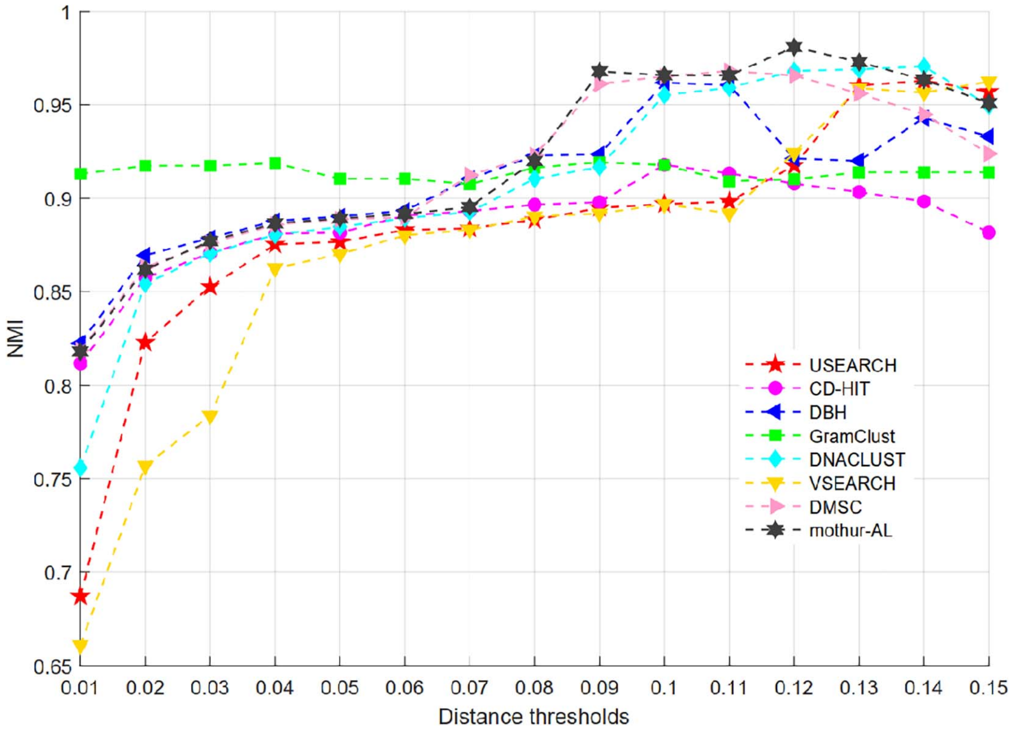


FIGURE 8 | NMI values of eight OTU picking methods at different clustering thresholds on the V4 dataset.

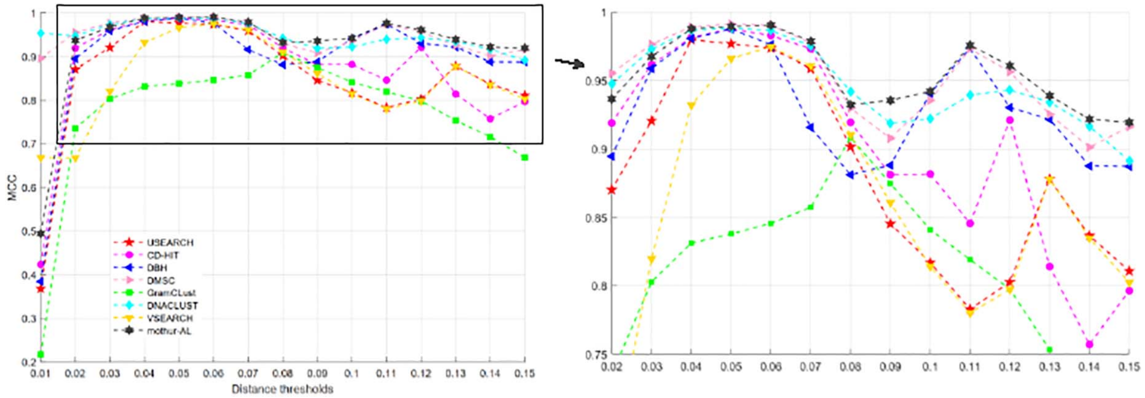


FIGURE 9 | MCC values of eight OTU picking methods with different clustering thresholds on the V4 dataset.

TABLE 4 | The average, SD, and maximum MCC values of seven OTU picking methods on V4 dataset.

	DMSC (0.05)	USEARCH (0.04)	VSEARCH (0.06)	DNACLUST (0.05)	DBH (0.05)	GramClust (0.08)	CD-HIT (0.05)	mothur-AL (0.06)
Max.	0.9913	0.9797	0.9746	0.9884	0.9875	0.9083	0.9876	0.9904
Ave.	0.9480	0.8481	0.8444	0.9478	0.8938	0.7671	0.8697	0.9246
SD	0.0330	0.1438	0.0933	0.0283	0.1409	0.1593	0.1382	0.1175

The values in the parentheses are the clustering thresholds where each method achieves its peak MCC.

Computational Complexity Analysis

Finally, in order to evaluate the computational complexity (including running time and memory usage) of different OTU picking methods, we used one large volume sequence dataset (V35) processed by QIIME from the HMP official website[2], which covers V3–V5 hypervariable regions and contains ∼30.3

[2] https://www.hmpdacc.org/hmp/HMQCP/

million sequences with an average length of 528 bp. It is reported that with sequencing coverage or sequences increase, the probability of duplicate sequences will be observed (Schloss and Westcott, 2011). Thus, for relatively fair comparisons across different OTU picking algorithms, the unique sequences (∼19.8 million) of V35 were used to evaluate the computational complexity of the OTU picking methods. We only report the computational complexity of nine heuristic methods of CD-HIT, DBH, DMSC, DNACLUST, DySC, GramClust, Swarm, USEARCH, and VESARCH because mothur-AL and CROP are time consuming for large-scale datasets, and ESPRIT-Forest always returns a core dumped information. **Supplementary Figure S12A** depicts the execution time (wall time) of nine OTU picking algorithms with different sequence sizes ranging from 10^4 to 10^6. It can be seen that the speed of DMSC is lower than that of other clustering methods. The speed of DBH, USEARCH, DNACLUST, and CD-HIT is faster than other methods when the sequence number increases. **Supplementary Figure S12B** graphically describes the memory usage for each method. We can obverse that DMSC and VESARCH consume more memory than other clustering methods, while Swarm, DySC, GramClust, and CD-HIT need less memory usage than other methods.

CONCLUSION AND PERSPECTIVES

With the development of high-throughput sequencing technologies, it has become convenient and cost efficient to thoroughly profile the microbial community composition and diversity in various environmental habitats (Deshpande et al., 2018; Escalona et al., 2018; Rodriguez-R et al., 2018; Fritz et al., 2019; Huang et al., 2021). Millions of sequencing data can be generated, and how to utilize this enormous sequence resource has become a critical concern for microbial ecologists (Szalkai and Grolmusz, 2018; Qu et al., 2019b). One particular challenge is the OTU picking in amplicon sequence analysis. Luckily, this challenge can be directly addressed by sequence clustering that attempts to group similar sequences (De Vrieze et al., 2018; Edgar, 2018). Therefore, numerous clustering methods have been proposed to help to unlock the great wealth contained in sequence datasets, but none of the methods notably outperforms all the others, and how to choose an appropriate method has become a challenge for inexperienced users. A lot of time and resources can be wasted in selecting clustering tools and analyzing the clustering results. In this review, we introduced the recent advance of clustering methods, which mainly focuses on three aspects: (i) the principles of existing clustering algorithms, (ii) benchmark dataset construction for OTU picking and evaluation metrics, and (iii) the performance of different methods with various similarity/distance thresholds on benchmark datasets. From the scope of clustering algorithms, we introduced the key clustering procedures for each category, such as hierarchical clustering methods, heuristic clustering methods, model-based methods, and network-based methods. From the scope of benchmark dataset construction and evaluation metrics, we introduced how to construct the ground-truth information for real-life 16S rRNA sequence datasets, presenting different criteria to evaluate clustering methods.

We compared the performance of the existing 12 state-of-art OTU picking methods of CD-HIT, USEARCH, DNACLUST, Swarm, VSEARCH, DBH, DMSC, DySC, ESPRIT-Forest, GramClust, mothur-AL, and CROP. It is found that the performance of most methods with different distance thresholds shows similar clustering results in terms of NMI. DMSC, DNACLUST, and USEARCH achieved the NMI peak values on the simulated dataset, V4 dataset, and full-length 16S rRNA dataset, respectively. In terms of MCC, mothur-AL achieved better clustering results on simulated dataset, DMSC had better clustering results for V4 datasets, and DBH obtained better clustering results on the full-length 16S rRNA dataset. Although numerous OTU picking methods have been proposed, mothur still is a competitive tool for amplicon sequence analysis. Concomitant with the large number of sequences produced by high-throughput technologies, four future directions to design the OTU picking algorithms should be paid attention to. One direction is to design the powerful clustering methods for huge sequences with longer sequence length. A striking challenge brought by the advent of sequencing technology is the rapid growth of sequence length. Several third-generation sequencing technologies (e.g., PacBio, Nanopore) (Rhoads and Au, 2015; Han et al., 2018; Ono et al., 2020) claim to have a long read length of 10∼100 kbp, which can cover the whole region of 16S rRNA gen (Wagner et al., 2016; Pootakham et al., 2017; Earl et al., 2018). Therefore, OTU picking methods for longer sequences will be in high demand. Another is clustering stability. From the comparison results in terms of MCC, we can see that the MCC curve of each method varies a lot with the distance threshold changes. The MCC curve should be a straight line for a stable clustering method, that is, given different distance thresholds, the OTU picking method should cluster sequences within the distance threshold into one group and the sequences beyond the distance threshold into different groups. The third is the integration of new clustering algorithms to the popular sequence analysis platforms or pipelines, such as mothur and QIIME2. When an excellent clustering algorithm was developed, developers should let their algorithm be expandable or easy to be applied into the platforms, so that the clustering results or outputs of a new method can be directly used as the input of relative commands in platforms, or the outputs from the platforms can be directly fed into the new method. This will be very convenient for users to adopt new clustering algorithms in the platform. The last direction is how to handle sequencing errors (Ma et al., 2019). Most existing OTU picking methods are just designed for sequence clustering, while the sequences generated by the sequencing platform will inevitably contain sequencing errors (Gaspar, 2018). Removing or reducing the sequencing errors will improve the accuracy of describing the microbial community. Although some error-correction (denoising) methods, such as DATA2 (Callahan et al., 2016), UNOISE (Edgar, 2016), Deblur (Amir et al., 2017), and SeekDeep (Hathaway et al., 2017), have been developed, how to combine these error-correction methods with OTU picking methods needs attention.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

Z-GW performed all the procedures using the clustering software, analyzed the clustering results, and wrote the manuscript. X-DZ downloaded the source codes and installed the software of all the clustering methods. MC and FL participated in the experimental studies and collected the benchmark datasets. YQ helped in improving the manuscript. S-WZ conceived the overall study, and reviewed and revised the manuscript. All authors read, edited, and approved the final manuscript.

FUNDING

This paper was supported by the National Natural Science Foundation of China (Nos. 61873202, 61473232, and 91430111) and the Scientific Research Program Funded by the Shaanxi Provincial Education Department (recipient: Z-GW, 2021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.644012/full#supplementary-material

REFERENCES

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., and Donati, C. (2015). MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Sci. Rep.* 5:9743.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., et al. (2019). A new genomic blueprint of the human gut microbiota. *Nature* 568, 499–504. doi: 10.1038/s41586-019-0965-1

Amir, A., Mcdonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16.

Barriuso, J., Valverde, J. R., and Mellado, R. P. (2011). Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *Bmc Bioinform.* 12:473. doi: 10.1186/1471-2105-12-473

Bazin, A., Debroas, D., and Mephu Nguifo, E. (2018). A de novo robust clustering approach for amplicon-based sequence data. *J. Comput. Biol.* 26, 618–624. doi: 10.1089/cmb.2018.0170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Cai, Y., and Sun, Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* 39:e95. doi: 10.1093/nar/gkr349

Cai, Y., Wei, Z., Jin, Y., Yang, Y., Mai, V., Qi, M., et al. (2017). ESPRIT-Forest: parallel clustering of massive amplicon sequence data in subquadratic time. *PLoS Comput. Biol.* 13:e1005518. doi: 10.1371/journal.pcbi.1005518

Callahan, B. J., Mcmurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581. doi: 10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336.

Chen, S. Y., Deng, F., Huang, Y., Jia, X., Liu, Y. P., and Lai, S. J. (2016). bioOTU: an improved method for simultaneous taxonomic assignments and operational taxonomic units clustering of 16s rRNA gene sequences. *J. Comput. Biol.* 23, 229–238. doi: 10.1089/cmb.2015.0214

Chen, W., Cheng, Y., Zhang, C., Zhang, S., and Zhao, H. (2013a). MSClust: a multi-seeds based clustering algorithm for microbiome profiling using 16S rRNA sequence. *J. Microbiol. Methods* 94, 347–355. doi: 10.1016/j.mimet.2013.07.004

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013b). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One* 8:e70837. doi: 10.1371/journal.pone.0070837

Cheng, L., Walker, A. W., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Res.* 40, 5240–5249. doi: 10.1093/nar/gks227

Chong, J., Liu, P., Zhou, G., and Xia, J. J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* 15, 799–821. doi: 10.1038/s41596-019-0264-1

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acid Res.* 37, D141–D145.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2013). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642.

De Filippis, F., Parente, E., Zotta, T., and Ercolini, D. (2018). A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. *Int. J. Food Microbiol.* 265, 9–17. doi: 10.1016/j.ijfoodmicro.2017.10.028

De Vrieze, J., Pinto, A. J., Sloan, W. T., and Ijaz, U. Z. (2018). The active microbial community more accurately reflects the anaerobic digestion process: 16S rRNA (gene) sequencing as a predictive tool. *Microbiome* 6:63.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/aem.03006-05

Deshpande, A., Lang, W., Mcdowell, T., Sivakumar, S., Zhang, J., Wang, J., et al. (2018). Strategies for identification of somatic variants using the Ion Torrent deep targeted sequencing platform. *BMC Bioinform.* 19:5. doi: 10.1186/s12859-017-1991-3

Earl, J. P., Adappa, N. D., Krol, J., Bhat, A. S., Balashov, S., Ehrlich, R. L., et al. (2018). Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome* 6:190.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10:996. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* [Preprint]. 081257.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Escalona, M., Rocha, S., and Posada, D. J. (2018). NGSphy: phylogenomic simulation of next-generation sequencing data. *Bioinformatics* 34, 2506–2507. doi: 10.1093/bioinformatics/bty146

Forster, D., Dunthorn, M., Stoeck, T., and Mahé, F. (2016). Comparison of three clustering approaches for detecting novel environmental microbial diversity. *PeerJ* 4:e1692. doi: 10.7717/peerj.1692

Franzén, O., Hu, J., Bao, X., Itzkowitz, S. H., Peter, I., and Bashir, A. (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome* 3:43.

Fritz, A., Hofmann, P., Majda, S., Dröge, J., Fiedler, J., Lesker, T. R., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 7, 1–12.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6:17.

Gao, F. (2018). Recent developments of software and database in microbial genomics and functional genomics. *Brief. Bioinform*. 20, 732–734. doi: 10.1093/bib/bby013

Gaspar, J. M. (2018). NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinform.* 19:536.

Gentile, C. L., and Weir, T. L. (2018). The gut microbiota at the intersection of diet and human health. *Science* 362, 776–780. doi: 10.1126/science.aau5812

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinform.* 12:271. doi: 10.1186/1471-2105-12-271

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol.* 12:69.

Han, R., Li, Y., Wang, S., Gao, X., Bi, C., and Li, M. (2018). DeepSimulator: a deep simulator for Nanopore sequencing. *Bioinformatics* 34, 2899–2908. doi: 10.1093/bioinformatics/bty223

Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27, 611–618. doi: 10.1093/bioinformatics/btq725

Hathaway, N. J., Parobek, C. M., Juliano, J. J., and Bailey, J. A. (2017). SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res.* 46:e21. doi: 10.1093/nar/gkx1201

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Huang, Y., Yuan, K., Tang, M., Yue, J., Bao, L., Wu, S., et al. (2021). Melatonin inhibiting the survival of human gastric cancer cells under ER stress involving autophagy and Ras-Raf-MAPK signalling. *J. Cell. Mol. Med.* 25, 1480–1492. doi: 10.1111/jcmm.16237

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Integrative HMP (iHMP) Research Network Consortium, (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16:276. doi: 10.1016/j.chom.2014.08.014

Jääskinen, V., Parkkinen, V., Cheng, L., and Corander, J. (2014). Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Stat. Applic. Genet. Mol. Biol.* 13, 105–121.

Jackson, M. A., Bell, J. T., Spector, T. D., and Steves, C. J. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ* 4:e2341. doi: 10.7717/peerj.2341

Jin, Y., and Bi, Z. (2018). "Power load curve clustering algorithm using fast dynamic time warping and affinity propagation," in *Proceedings of the 2018 5th International Conference on Systems and Informatics (ICSAI)*, (Nanjing: IEEE), 1132–1137.

Jo, H. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/mmbr.68.4.669-685.2004

Kellenberger, E. (2001). Exploring the unknown. *EMBO Rep.* 2, 5–7.

Kim, M., Lee, K.-H., Yoon, S.-W., Kim, B.-S., Chun, J., and Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform.* 11:102. doi: 10.5808/gi.2013.11.3.102

Koslicki, D., Foucart, S., and Rosen, G. (2013). Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics* 29, 2096–2102. doi: 10.1093/bioinformatics/btt336

Lapierre, N., Ju, C. J., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* 166, 74–82. doi: 10.1016/j.ymeth.2019.03.003

Laudadio, I., Fulci, V., Stronati, L., and Carissimi, C. (2019). Next-generation metagenomics: methodological challenges and opportunities. *OMICS* 23, 327–333. doi: 10.1089/omi.2019.0073

Lawley, B., and Tannock, G. W. (2017). "Analysis of 16S rRNA gene amplicon sequences using the QIIME software package," in *Oral Biology*, eds G. Seymour, M. Cullinan, and N. Heng, (New York, NY: Humana Press), 153–163. doi: 10.1007/978-1-4939-6685-1_9

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.* 13, 656–668. doi: 10.1093/bib/bbs035

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, F., Zhang, S.-W., Wei, Z.-G., Chen, W., and Zhou, C. (2014). Mining seasonal marine microbial pattern with greedy heuristic clustering and symmetrical nonnegative matrix factorization. *BioMed Res. Int.* 2014:189590.

Liu, Z., Liu, Y., Dezert, J., and Cuzzolin, F. (2019). Evidence combination based on credal belief redistribution for pattern classification. *IEEE Trans. Fuzzy Syst.* 28, 618–631. doi: 10.1109/tfuzz.2019.2911915

Liu, Z., Pan, Q., Dezert, J., Han, J.-W., and He, Y. (2018). Classifier fusion with contextual reliability evaluation. *IEEE Trans. Cybernet.* 48, 1605–1618. doi: 10.1109/tcyb.2017.2710205

Liu, Z., Pan, Q., Dezert, J., and Martin, A. (2017). Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Trans. Fuzzy Syst.* 26, 1217–1230. doi: 10.1109/tfuzz.2017.2718483

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20, 1–15.

Magoè, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Mahé, F., Rognes, T., Quince, C., De Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593. doi: 10.7717/peerj.593

Matias Rodrigues, J. F., Schmidt, T. S., Tackmann, J., and Von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 33, 3808–3810. doi: 10.1093/bioinformatics/btx517

Matias Rodrigues, J. F., and von Mering, C. (2013). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 30, 287–288. doi: 10.1093/bioinformatics/btt657

Mercier, C., Boyer, F., Bonin, A., and Coissac, E. (2013). "SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences," in *Programs and Abstracts of the SeqBio 2013 Workshop. Abstract*, (Citeseer), 27–29. Available online at: https://git.metabarcoding.org/obitools/sumatra/wikis/home

Meyer, F., Bremges, A., Belmann, P., Janssen, S., Mchardy, A. C., and Koslicki, D. (2019). Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* 20:51.

Namiki, Y., Ishida, T., and Akiyama, Y. (2013). Acceleration of sequence clustering using longest common subsequence filtering. *BMC Bioinform.* 14:S7.

Nguyen, T.-D., Schmidt, B., Zheng, Z., and Kwoh, C.-K. (2015). Efficient and accurate OTU clustering with GPU-based sequence alignment and dynamic dendrogram cutting. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 1060–1073. doi: 10.1109/tcbb.2015.2407574

Niu, B., Fu, L., Sun, S., and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinform.* 11:187. doi: 10.1186/1471-2105-11-187

Niu, B., Zhu, Z., Fu, L., Wu, S., and Li, W. (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 27, 1704–1705. doi: 10.1093/bioinformatics/btr252

Ono, Y., Asai, K., and Hamada, M. (2020). PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics* btaa835. doi: 10.1093/bioinformatics/btaa835

Pagni, M., Niculita-Hirzel, H., Pellissier, L., Dubuis, A., Xenarios, I., Guisan, A., et al. (2013). Density-based hierarchical clustering of pyro-sequences on a large scale—the case of fungal ITS1. *Bioinformatics* 29, 1268–1274. doi: 10.1093/bioinformatics/btt149

Peterson, C. T., Sharma, V., Iablokov, S. N., Albayrak, L., Khanipov, K., Uchitel, S., et al. (2019). 16S rRNA gene profiling and genome reconstruction reveal community metabolic interactions and prebiotic potential of medicinal herbs used in neurodegenerative disease and as nootropics. *PLoS One* 14:e0213869. doi: 10.1371/journal.pone.0213869

Peterson, J., Garges, S., Giovanni, M., Mcinnes, P., and Guyer, M. (2009). *The NIH Human Microbiome Project*. Hoboken, NJ: John Wiley & Sons, Ltd.

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Sonthirod, C., Naktang, C., et al. (2017). High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci. Rep.* 7:2774.

Qian, Y., Zhang, C., Wei, Z., Liu, F., Yao, C., and Zheng, Z. J. E. (2020). The optimal oscillation mode in excitable small-world networks. *EPL* 131:38002. doi: 10.1209/0295-5075/131/38002

Qian, Y., Zhang, G., Wang, Y., Yao, C., and Zheng, Z. (2019). Winfree loop sustained oscillation in two-dimensional excitable lattices: Prediction and realization. *Chaos Interdis. J. Nonlinear Sci.* 29:073106. doi: 10.1063/1.5085644

Qu, K., Gao, F., Guo, F., and Zou, Q. (2019a). Taxonomy dimension reduction for colorectal cancer prediction. *Comput. Biol. Chem.* 83:107160. doi: 10.1016/j.compbiolchem.2019.107160

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019b). Application of machine learning in microbiology. *Front. Microbiol.* 10:827. doi: 10.3389/fmicb.2019.00827

Rambaut, A., and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13, 235–238. doi: 10.1093/bioinformatics/13.3.235

Rasheed, Z., Rangwala, H., and Barbará, D. (2013). 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. *BMC Syst. Biol.* 7:S11.

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi: 10.1146/annurev.genet.38.072902.091216

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018). Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* 3:e00039-18.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Russell, D. J., Way, S. F., Benson, A. K., and Sayood, K. (2010). A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinform.* 11:601.

Sanli, K., Bengtsson-Palme, J., Nilsson, R. H., Kristiansson, E., Alm Rosenblad, M., Blanck, H., et al. (2015). Metagenomic sequencing of marine periphyton: taxonomic and functional insights into biofilm communities. *Front. Microbiol.* 6:1192.

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* 6:e1000844. doi: 10.1371/journal.pcbi.1000844

Schloss, P. D. (2016). Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. *mSystems* 1:e00027-16.

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71, 1501–1506. doi: 10.1128/aem.71.3.1501-1506.2005

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 77, 3219–3226. doi: 10.1128/aem.02810-10

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/aem.01541-09

Seguritan, V., and Rohwer, F. (2001). FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinform.* 2:9. doi: 10.1186/1471-2105-2-9

Shah, M. S., Desantis, T. Z., Weinmaier, T., Mcmurdie, P. J., Cope, J. L., Altrichter, A., et al. (2018). Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* 67, 882–891. doi: 10.1136/gutjnl-2016-313189

Stackebrandt, E., and Goebel, B. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849. doi: 10.1099/00207713-44-4-846

Streit, W. R., and Schmitz, R. A. (2004). Metagenomics–the key to the uncultured microbes. *Curr. Opin. Microbiol.* 7, 492–498. doi: 10.1016/j.mib.2004.08.002

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., et al. (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.* 13, 107–121. doi: 10.1093/bib/bbr009

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., Mckendree, W., et al. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* 37:e76. doi: 10.1093/nar/gkp285

Szalkai, B., and Grolmusz, V. J. B. (2018). SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification. *Bioinformatics* 34, 2487–2489. doi: 10.1093/bioinformatics/bty116

Thaiss, C. A. (2018). Microbiome dynamics in obesity. *Science* 362, 903–904. doi: 10.1126/science.aav6870

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449, 804–810.

Větrovský, T., Baldrian, P., and Morais, D. (2018). SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. *Bioinformatics* 34, 2292–2294. doi: 10.1093/bioinformatics/bty071

Wagner, J., Coupland, P., Browne, H. P., Lawley, T. D., Francis, S. C., and Parkhill, J. J. B. M. (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.* 16:274.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07

Wang, X., Yao, J., Sun, Y., and Mai, V. (2013). M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinform.* 14:43. doi: 10.1186/1471-2105-14-43

Ward, D. M., Weller, R., and Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345:63. doi: 10.1038/345063a0

Wei, Z., and Zhang, S.-W. (2019). DMSC: a dynamic multi-seeds method for clustering 16S rRNA sequences into OTUs. *Front. Microbiol.* 10:428.

Wei, Z.-G., and Zhang, S.-W. (2015). MtHc: a motif-based hierarchical method for clustering massive 16S rRNA sequences into OTUs. *Mol. Biosyst.* 11, 1907–1913. doi: 10.1039/c5mb00089k

Wei, Z.-G., and Zhang, S.-W. (2017). DBH: A de Bruijn graph-based heuristic method for clustering large-scale 16S rRNA sequences into OTUs. *J. Theor. Biol.* 425, 80–87. doi: 10.1016/j.jtbi.2017.04.019

Wei, Z.-G., and Zhang, S.-W. (2018). NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. *BMC Bioinform.* 19:177.

Wei, Z.-G., Zhang, S.-W., and Jing, F. (2016). Exploring the interaction patterns among taxa and environments from marine metagenomic data. *Quantitative Biol.* 4, 84–91. doi: 10.1007/s40484-016-0071-4

Wei, Z. G., Zhang, S. W., and Liu, F. J. (2020). smsMap: mapping single molecule sequencing reads by locating the alignment starting positions. *BMC Bioinform.* 21:341.

Wei, Z. G., Zhang, S. W., and Zhang, Y. Z. (2017). DMclust, a density-based Modularity method for accurate OTU picking of 16S rRNA sequences. *Mol. Inform.* 36:1600059. doi: 10.1002/minf.201600059

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., et al. (2020). Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microbiome* 15, 1–12.

Westcott, S. L., and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. doi: 10.7717/peerj.1487

Westcott, S. L., and Schloss, P. D. (2017). OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2:e00073-17.

Whelan, F. J., and Surette, M. G. (2017). A comprehensive evaluation of the sl1p pipeline for 16S rRNA gene sequencing analysis. *Microbiome* 5, 1–13.

Woloszynek, S., Zhao, Z., Chen, J., and Rosen, G. L. (2019). 16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses. *PLoS Comput. Biol.* 15:e1006721. doi: 10.1371/journal.pcbi.1006721

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12:444.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5:e16. doi: 10.1371/journal.pbio.0050016

Zhang, S.-W., and Wei, Z.-G. (2015). Some remarks on prediction of protein-protein interaction with machine learning. *Med. Chem.* 11, 254–264. doi: 10.2174/1573406411666141230095838

Zhang, S. W., Wei, Z. G., Zhou, C., Zhang, Y. C., and Zhang, T. H. (2013). "Exploring the interaction patterns in seasonal marine microbial communities with network analysis," in *Proceedings of the 2013 9th International Conference on Systems Biology*, Huangshan, 63–68.

Zheng, Z., Kramer, S., and Schmidt, B. (2012). DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics* 28, 2182–2183. doi: 10.1093/bioinformatics/bts355

Zhu, Z., Ren, J., Michail, S., and Sun, F. (2019). MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biol.* 20, 1–13.

Zorita, E. V., Cusco, P., and Filion, G. J. (2015). Starcode: sequence clustering based on all-pairs search. *Bioinformatics* 31, 1913–1919. doi: 10.1093/bioinformatics/btv053

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

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.

An Inductive Logistic Matrix Factorization Model for Predicting Drug-Metabolite Association With Vicus Regularization

Yuanyuan Ma[1], Lifang Liu[2], Qianjun Chen[3] and Yingjun Ma[4]*

[1] School of Computer and Information Engineering, Anyang Normal University, Anyang, China, [2] School of Education, Anyang Normal University, Anyang, China, [3] School of Computer, Central China Normal University, Wuhan, China, [4] School of Applied Mathematics, Xiamen University of Technology, Xiamen, China

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Jin-Xing Liu,
Qufu Normal University, China
Wen Zhang,
Huazhong Agricultural University, China

Correspondence:
Yuanyuan Ma
chonghua_1983@126.com

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

Received: *07 January 2021*
Accepted: *08 March 2021*
Published: *01 April 2021*

Citation:
Ma Y, Liu L, Chen Q and Ma Y (2021) An Inductive Logistic Matrix Factorization Model for Predicting Drug-Metabolite Association With Vicus Regularization. Front. Microbiol. 12:650366. doi: 10.3389/fmicb.2021.650366

Metabolites are closely related to human disease. The interaction between metabolites and drugs has drawn increasing attention in the field of pharmacomicrobiomics. However, only a small portion of the drug-metabolite interactions were experimentally observed due to the fact that experimental validation is labor-intensive, costly, and time-consuming. Although a few computational approaches have been proposed to predict latent associations for various bipartite networks, such as miRNA-disease, drug-target interaction networks, and so on, to our best knowledge the associations between drugs and metabolites have not been reported on a large scale. In this study, we propose a novel algorithm, namely inductive logistic matrix factorization (ILMF) to predict the latent associations between drugs and metabolites. Specifically, the proposed ILMF integrates drug–drug interaction, metabolite–metabolite interaction, and drug-metabolite interaction into this framework, to model the probability that a drug would interact with a metabolite. Moreover, we exploit inductive matrix completion to guide the learning of projection matrices U and V that depend on the low-dimensional feature representation matrices of drugs and metabolites: F^m and F^d. These two matrices can be obtained by fusing multiple data sources. Thus, F^dU and F^mV can be viewed as drug-specific and metabolite-specific latent representations, different from classical LMF. Furthermore, we utilize the Vicus spectral matrix that reveals the refined local geometrical structure inherent in the original data to encode the relationships between drugs and metabolites. Extensive experiments are conducted on a manually curated "DrugMetaboliteAtlas" dataset. The experimental results show that ILMF can achieve competitive performance compared with other state-of-the-art approaches, which demonstrates its effectiveness in predicting potential drug-metabolite associations.

Keywords: logistic matrix factorization, drug-metabolite association, Vicus matrix, human metabolites, graph regularization

INTRODUCTION

With the development of metabolomics technology, more and more metabolites have been identified. This progress provides unprecedented opportunities to obtain new insights into the effects of drugs on metabolites. Recently, Liu et al. (2020) integrated epidemiologic, pharmacologic, genetic, and gut microbiome data to analyze the relationships between drugs and metabolites, which provided a trail for targeted experimental pharmaceutical research to improve drug safety and efficacy. Exploring the potential drug-metabolite associations is also a novel route towards pharmacomicrobiomics and precision medicine. Doestzada et al. (2018) reviewed the complex interactions between host, intestinal microorganisms and drugs, and thought that pharmacomicrobiomics would provide an important foundation for personalized medicine and precision medicine. The earliest report about interactions between drugs and metabolites can be dated back to the 1930s with the discovery of sulphanilamide (Fuller, 1937). The activity of prontosil is due to the transformation of microbial azoreductases and the liberation of sulphanilamide. In addition, microbial metabolites can also inactivate drugs, such as digoxin. A study on *Eggerthella lenta* strains in 2013 (Haiser et al., 2013) found that these strains carried a two-gene cardiac glycoside reductase (cgr) operon that was transcriptionally activated by digoxin (Doestzada et al., 2018), and thus resulted in the inactivation of the drug in cardiovascular treatment.

Identifying drug-metabolite associations not only provides deep insights into understanding complex interaction mechanisms among them, but it can also benefit the screening of chemical compounds for drug development and improve microbe related therapy. The complex relationship between drugs, metabolites, and microbes has attracted extensive attention. However, conventional wet-lab research for verifying drug-metabolite interactions is generally labor-intensive, costly, and time-consuming. Computational approaches are a viable alternative. Shang et al. (2014) found that metabolites in the same pathway were usually associated with the similar or same disease. Based on this fact, they proposed a metabolite pathway-based random walk algorithm to prioritize the candidate disease metabolites (Shang et al., 2014). Yao et al. (2015) presented an approach based on global distance similarity to predict and prioritize disease related metabolites. Ma et al. (2020b) integrated multiple diseases and metabolite similarity networks to predict the potential associations between metabolites and diseases. Long and Luo (2020) used multi-source biomedical data to construct a three-level heterogeneous network and designed a novel network embedding representation framework to identify microbe-drug associations. Specifically, Long et al. (2020) exploited the conditional random field, graph convolutional network and a random walk with restart (RWR) to learn the latent feature representations of drugs and microbes, identifying some potential drug-microbe associations.

Although these studies have obtained some valuable results, there are two main limitations to the existing drug-microbe or metabolite-disease association mining approaches. Firstly, the accuracy of these methods is still unsatisfactory due to a lack of sufficient prior information for drugs, microbes, and diseases. Secondly, the local geometrical structure of nodes is important in the task of dimensionality deduction and data representation, which decides the effectiveness and efficiency of algorithms to a large extent. The algorithms mentioned above did not consider the local spectral information that resides in the original data, meaning their performances are not ideal.

In this study, we propose a novel computational approach, named inductive logistic matrix factorization (ILMF), to analyze latent drug-metabolite associations. ILMF integrates the advantages of logistic matrix factorization (LMF; Johnson, 2014; Liu et al., 2016) and inductive matrix completion (Natarajan and Dhillon, 2014; Chen et al., 2018) to learn low-dimensional embedding of drugs and metabolites, and predict the final interaction probabilities based on the two low-dimensional representation of drugs and metabolites. Specifically, ILMF first learns the latent representation of drugs and metabolites via clusDCA (Cho et al., 2015; Wang et al., 2015), which runs RWR on each node in each interaction network (e.g., metabolite–metabolite interaction network or similarity network) to compute "the diffusion state" of each point, and then utilizes a singular value decomposition (SVD)-based approach to obtain the consensus low-dimensional matrix representation for metabolites and drugs Z^m and Z^d, respectively. Secondly, based on Z^m and Z^d, ILMF exploits LMF to learn two projection matrices U and V, respectively, so that Z^mV and Z^dU have the same semantic space. Finally, a logistic function is used to predict the probability that a drug would interact with a metabolite in the same way that LMF does. Nevertheless, in contrast to LMF, ILMF captures the topological properties of nodes (i.e., drugs or metabolites) and takes advantage of the idea of inductive matrix completion (Luo et al., 2017) to generate the optimal projection of drugs and metabolites. In addition, ILMF also exploits the local spectral Vicus matrices (Wang B. et al., 2017) of drugs and metabolites to reveal the refined local geometrical structure inherent in drug–drug interaction network and metabolite–metabolite interaction network. An illustrative example of this pipeline is given in **Figure 1**, followed by a more detailed description of ILMF in section "Materials and Methods."

The contributions of this article are summarized as follows:

1. We propose a novel LMF-based framework, named ILMF, to predict drug-metabolite associations by integrating multiple biological networks. To the best of our knowledge, this is the first work to predict the latent drug-metabolite associations.
2. ILMF combines the advantages of inductive matrix completion and the local spectral Vicus matrix of each interaction network into this framework, and captures the optimal low-dimensional representation of drugs and metabolites.
3. We have manually curated a drug-metabolite association dataset ("DrugMetaboliteAtlas") by retrieving relevant literature. This benchmark dataset can be used to evaluate the performance of various association prediction algorithms, which facilitates future research in drug-metabolite association prediction tasks.

The comprehensive experiments show that the proposed ILMF algorithm outperforms several state-of-the-art methods on the curated "DrugMetaboliteAtlas" dataset. In addition, the prediction ability of ILMF has also been confirmed by retrieving the latest published literature or information from databases.

MATERIALS AND METHODS

Materials

The "DrugMetaboliteAtlas" dataset was downloaded from the BBRMI-NL website[1] (Liu et al., 2020). It contains 1071 interactions from 87 commonly prescribed drugs and 150 clinically relevant metabolites. After removing drugs lacking significantly relevant metabolite associations, 42 drugs were reserved. In addition, we also manually curated the correlations between drug categories and the correlations between metabolites in the Rotterdam study (Liu et al., 2020).

Metabolite-microbe associations and metabolite-pathway associations were also downloaded from literature (Kurilshikov et al., 2019). The metabolite similarities from each type of association were computed based on the Gaussian interaction profile kernel (He et al., 2018; Ma et al., 2020b). After that, clusDCA (Wang et al., 2015) was used to fuse multiple drug–drug interaction networks and multiple metabolite–metabolite interaction networks. Simultaneously, the optimal low-dimensional matrix representations of metabolites and drugs F^m, F^d can also be obtained from this fusing process. Then, the local Vicus spectral matrices of metabolites and drugs V^{irm}, V^{ird} were computed based on the optimal low-dimensional matrix representations of metabolites and drugs F^m and F^d, respectively. Finally, the low-dimensional feature matrices of drugs and metabolites F^m and F^d, the local spectral matrices Vir^m and Vir^d were used as input of the proposed ILMF algorithm.

Problem Formalization

In this article, the set of drugs is denoted by $D = \{d_i\}_{i=1}^{n}$, and the set of metabolites is denoted by $M = \{m_j\}_{j=1}^{m}$, where, n and m are the number of drugs and metabolites, respectively. The known drug-metabolite interactions are represented as a $n \times m$ binary matrix $Y \in R^{n\times m}$, where $y_{ij} = 1$ if a drug d_i has been observed to interact with a metabolite m_j; otherwise $y_{ij} = 0$.

This study aimed to solve the problem of predicting the interaction probability of a drug-metabolite interaction pair, and subsequently rank the candidate drug-metabolite pairs based on these probabilities in descending order. Thus, the top-ranked pairs can be viewed as the most relevant interactions.

Metabolite–Metabolite Similarity

There are four metabolite related data sources: metabolite–metabolite correlation matrix Cor_m, metabolite-microbial species association matrix MM, metabolite-pathway association matrix MP, and drug-metabolite interaction matrix Y. Cor_m is obtained from literature (Liu et al., 2020); MM and MP are collected from literature (Kurilshikov et al., 2019).

For drug-metabolite association matrix Y, we use the Gaussian interaction profile kernel (He et al., 2018) to compute the similarity between any two metabolites. Let the j-th column $y_{\cdot j}$ of Y denote the interaction profile between metabolite m_j and all drugs. For any two metabolites m_i and m_j, the similarity between them can be measured as:

$$K_{md} = \exp\left(-\gamma_m ||y_{\cdot i} - y_{\cdot j}||^2\right). \quad (1)$$

Where γ_m is a bandwidth parameter that needs to be normalized based on a new bandwidth parameter γ'_m :

$$\gamma_m = \gamma'_m \bigg/ \left(\frac{1}{m}\sum_{l=1}^{m} |y_{\cdot l}|^2\right). \quad (2)$$

Here, m is the number of metabolites. $|\cdot|$ denotes Frobenius norm. γ'_m is set to be 1 according to the previous study (Wang F. et al., 2017; He et al., 2018).

The Gaussian profile kernel similarity matrices K_{mm} and K_{mp} can also be computed based on metabolite-microbial species association matrix MM and metabolite-pathway association matrix MP, respectively.

Drug–Drug Similarity

There are two drug related data sources: drug–drug correlation matrix Cor_d and drug-metabolite interaction matrix Y. Cor_d, which were obtained from literature (Liu et al., 2020). Analogously, the Gaussian interaction profile kernel similarity matrix K_d between any two drugs can be computed in the same way.

After obtaining four metabolite–metabolite similarity matrices and two drug–drug similarity matrices derived from multiple data sources, we used clusDCA (Cho et al., 2015; Wang et al., 2015) to fuse these similarity matrices and finally acquire the optimal low-dimensional matrix representations of metabolite and drug features F^m and F^d, respectively.

Inductive Logistic Matrix Factorization

Logistic matrix factorization has been demonstrated to be effective in the prediction of drug-target interactions (Liu et al., 2016), metabolite-disease (Ma et al., 2020a), and personalized recommendations (Hu et al., 2008; Johnson, 2014; Liu et al., 2014). The main advantage of LMF is that it assigns higher levels of importance to the observed interaction pairs than unknown ones. In this study, we apply LMF for drug-metabolite interaction prediction. LMF maps drugs and metabolites into a shared low-dimensionality latent semantic space $r \ll \min(m, n)$. The interaction probability p_{ij} of a drug-metabolite pair (d_i, m_j) can be modeled as follows:

$$p_{ij} = \frac{\exp\left(w_i h'_j\right)}{1 + \exp\left(w_i h'_j\right)}. \quad (3)$$

Where $w_i \in R^{1\times r}$, $h_j \in R^{1\times r}$ are latent representations of drug d_i and metabolite m_j, respectively. For

[1] http://bbmri.researchlumc.nl/atlas/

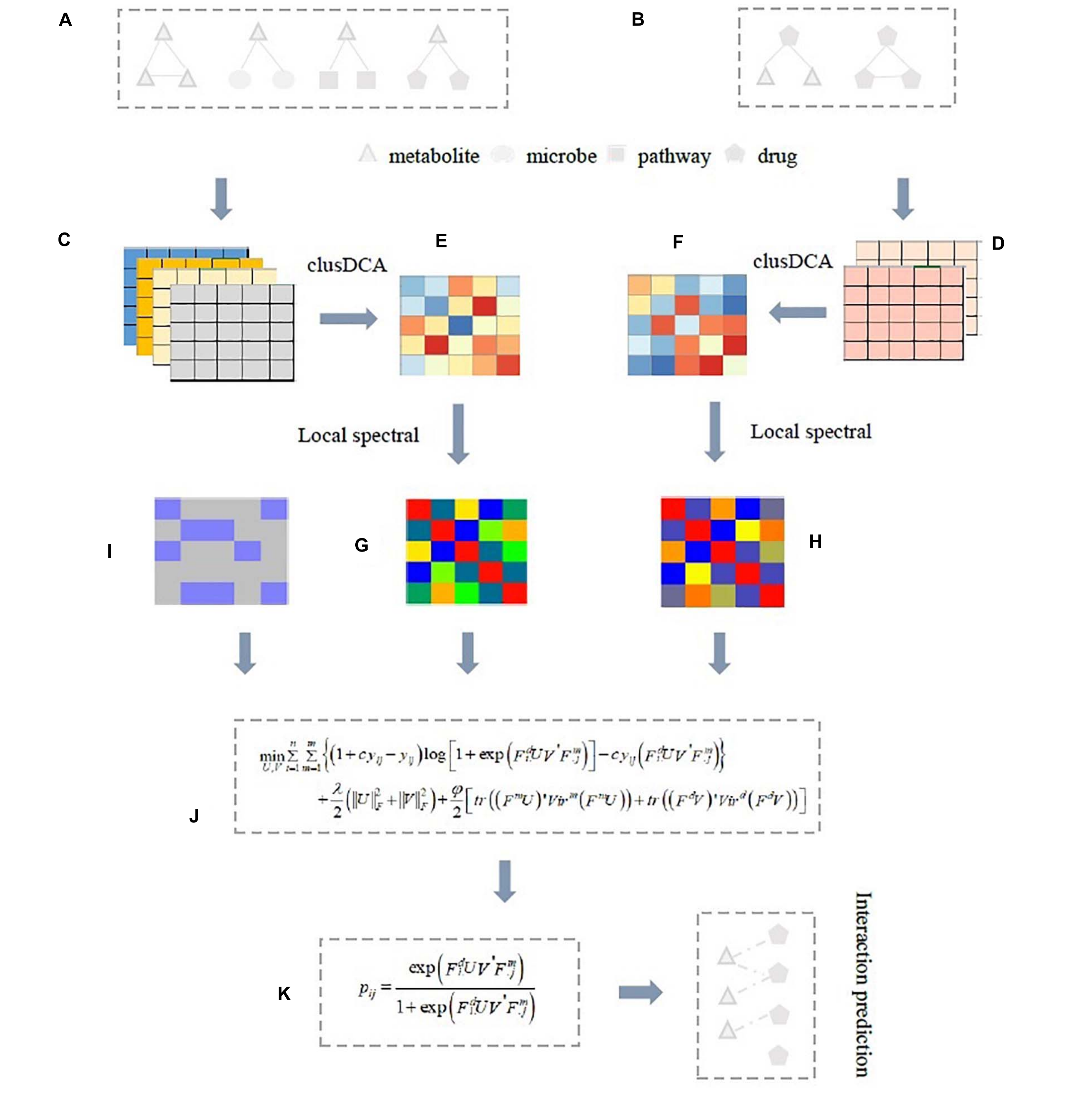


FIGURE 1 | Illustrative example of ILMF for predicting potential drug-metabolite associations. **(A)** Metabolite–metabolite, metabolite-drug, metabolite-microbe, metabolite-pathway association matrices, or correlation matrices; **(B)** Drug-metabolite, drug–drug association, or correlation matrices; **(C,D)** Based on Gaussian interaction profile kernel function, metabolite–metabolite similarity matrices, and drug–drug similarity matrices obtained from four metabolite association data and two drug association data, respectively; **(E)** The fused metabolite–metabolite similarity matrix by integrating four metabolite-related data with clusDCA; **(F)** The fused drug–drug similarity matrix by integrating two drug association data with clusDCA. Then, the local spectral matrix of metabolites **(G)** And the local spectral matrix of drugs **(H)** Can be obtained based on these two fused similarity matrices with Vicus; **(I)** The drug-metabolite association matrix; **(J)** The proposed ILMF model. Finally, ILMF outputs the predicted drug-metabolite interaction probability scores **(K)**. Here, a solid line indicates known associations, a dotted line indicates predicted drug-metabolite associations obtained from ILMF.

convenience, we further represent the latent vectors of all drugs and metabolites as matrix form $W \in R^{n\times r}$ and $H \in R^{m\times r}$, respectively.

The observed drug-metabolite interaction pairs are generally more reliable and important than the unknown interaction pairs. A higher level of importance was thus assigned to known

TABLE 1 | The pseudocode of the ILMF algorithm.

```
Input: The known association matrix $Y$; parameters $\lambda$, $\phi$, $c$, $K$
Output: The projection matrices, $U$ and $V$
1. Compute metabolite–metabolite similarity matrices $K_{md}$, $K_{mm}$, $K_{mp}$ according to Eqs 1 and 2, respectively; similarly, compute drug–drug similarity matrices $K_d$;
2. Compute the low-dimensional feature representational matrices of metabolites and drugs, $F^m$ and $F^d$ using clusDCA (Cho et al., 2015); computing Vicus spectral matrices of metabolites and drugs, $Vir^m$ and $Vir^d$;
3. Initialize $U$ and $V$ randomly;
4. For $t = 1, \ldots\ldots$, max_iter do
5. Update $U$ and $V$ according to AdaGrad algorithm
6. Until convergence conditions are satisfied
7. End for
8. Return $U$, $V$
```

TABLE 2 | The best performance of all methods on the "DrugMetaboliteAtlas" dataset.

	AUC	AUPR	F1
DTInet	0.7430	0.2176	0.2951
IMCMDA	0.7913	0.3655	0.4345
GRNMF	0.9272	0.5847	0.5767
ILMF⁻	0.9223	0.5429	0.5662
ILMF	0.9402	0.6303	0.6052

To ensure a fair comparison, the optimal parameters are selected from the ranges provided by these corresponding studies. For ILMF⁻ and ILMF, the above results are obtained when c = 2, ϕ =1, λ = 8, and $_{r=12}$.

interaction pairs than unknown ones. According to a previous study, we set the importance level to be $c\,(c \geq 1)$. Eq. 3 can be written as follows:

$$p\,(Y\,|U,V) = \prod_{1\leq i\leq n,\ 1\leq j\leq m,\ y_{ij}=1} \left[p_{ij}^{y_{ij}}\left(1-p_{ij}\right)^{(1-y_{ij})}\right]^{c} \times \prod_{1\leq i\leq n,\ 1\leq j\leq m,\ y_{ij}=0} \left[p_{ij}^{y_{ij}}\left(1-p_{ij}\right)^{(1-y_{ij})}\right]. \quad (4)$$

Here, c is the important level parameter used to control the weight assigned to the observed drug-metabolite pairs. In the next experiments, we empirically set it to two.

Inspired by the ideas of inductive matrix completion (Jain and Dhillon, 2013; Zeng et al., 2020) and generalized matrix factorization (GMF) (Zhang et al., 2020), we designed a novel ILMF framework, ILMF, to predict the latent interaction probabilities between drugs and metabolites. In particular, we used $F^d \in R^{n\times k_1}$ and $F^m \in R^{m\times k_2}$ derived from clusDCA (see section "Drug–Drug Similarity") to guide the learning process of projection matrices $U \in R^{k_1\times r}$ and $V \in R^{k_2\times r}$, so that the latent representations of metabolites and drugs $W = F^d U \in R^{n\times r}$ and $H = F^m V \in R^{m\times r}$ can carry compatible and complementary information from multiple data sources. Thus, in the ILMF model, Eq. 3 can be rewritten as follows:

$$p_{ij} = \frac{\exp\left(F_{i.}^{d} U V' F_{.j}^{m}\right)}{1+\exp\left(F_{i.}^{d} U V' F_{.j}^{m}\right)}. \quad (5)$$

Where $F_{i.}^{d}$ denotes the i-th row of F^d, $F_{.j}^{m}$ denotes the j-th column of F^m. By substituting Eq. 5 into Eq. 4, we estimate the projection matrices U and V by maximizing the above likelihood function (Eq. 3), which is equivalent to minimizing the negative logarithm of Eq. 3. Thus, the objective function of the proposed ILMF framework can be defined as:

$$\min_{U,V} \sum_{i=1}^{n}\sum_{m=1}^{m} \left(1 + c y_{ij} y_{ij} - y_{ij}\right) \log\left[1+\exp\left(F_{i.}^{d} U V' F_{.j}^{m}\right)\right] - c y_{ij}\left(F_{i.}^{d} U V' F_{.j}^{m}\right). \quad (6)$$

To avoid overfitting, the L$_2$ regularization is generally imposed on U and V. Thus, Eq. 6 becomes:

$$\min_{U,V} \sum_{i=1}^{n}\sum_{m=1}^{m} \left\{\left(1 + c y_{ij} - y_{ij}\right) \log\left[1+\exp\left(F_{i.}^{d} U V' F_{.j}^{m}\right)\right] - c y_{ij}\left(F_{i.}^{d} U V' F_{.j}^{m}\right)\right\} + \frac{\lambda}{2}\left(|U|_F^2 + |V|_F^2\right), \quad (7)$$

Where λ is a regularization parameter used to tradeoff the balance between reconstruction errors and smooth solutions.

Note that, for new drugs (metabolites) that do not have any known connections with metabolites (drugs), ILMF can still predict their potential associations, once we get their similarity network from other data sources. This is different from GMF (Zhang et al., 2020). In GMF, the neighborhood information of nodes was used to generate two feature matrices, and then they were adaptively updated at each iteration. In contrast, ILMF fuses multiple similarity networks to produce the low-dimensional matrix representations of metabolites and drugs.

Vicus Matrix

As demonstrated in literature (Wang B. et al., 2017), Vicus has many of the same properties as Laplacian. However, compared with Laplacian, Vicus can capture the local geometrical structure that resides within the original data well. The reason for using Vicus instead of Laplacian is that the local connection information from neighboring nodes makes the learned graph more robust to noise and helps to alleviate the influence of outliers.

Let $\{x_1, x_2, \ldots, x_n\}$ be the set of data points. Corresponding to x_i, v_i denotes the i-th vertex in a weighted network P, and $N\,(i)$ represents x_i 's neighborhood, not including x_i. Here, the neighborhood size of all nodes is consistent ($_{|N_i|=k,\ i=1,2,\ldots,n}$).

Based on the assumption that the cluster label of the i-th data point can be inferred from its nearest neighborhood $N\,(i)$, we first extract a subnetwork $P_i = (V_i E_i)$ such that V$_i$=$N\,(i) \bigcup x_i$. E_i represents the edges connecting all points in V_i. Using the label diffusion algorithm (Zhou et al., 2004), a virtual label indicator vector $c_{V_i}^{k}$ can be reconstructed as:

$$c_{v_i}^{k} = (1-\alpha)\,(I - \alpha S_i)^{-1} q_{v_i}^{k}, \quad 1 \leq k \leq C. \quad (8)$$

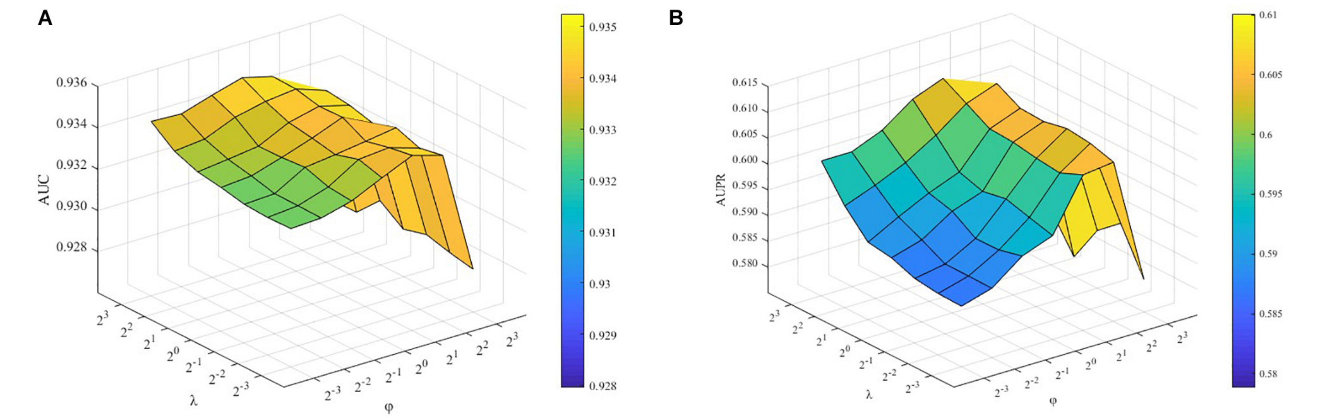


FIGURE 2 | Performance of ILMF on "DrugMetaboliteAtlas" dataset with different values of λ and ϕ. **(A)** AUC versus λ and ϕ; **(B)** AUPR versus λ and ϕ.

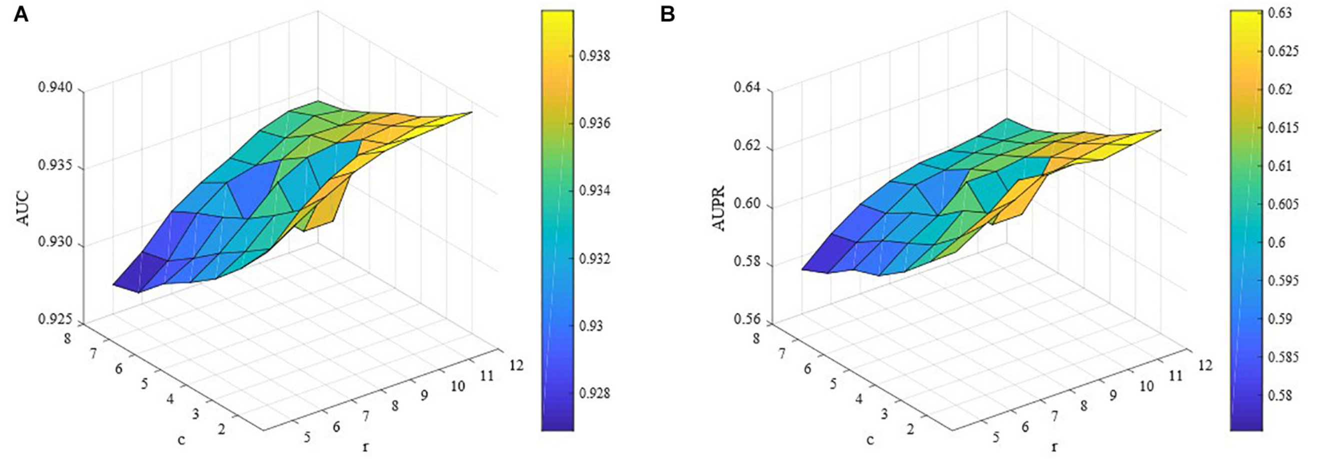


FIGURE 3 | Performance of ILMF on "DrugMetaboliteAtlas" dataset with different values of *c* and *r*. **(A)** AUC versus *c* and *r*; **(B)** AUPR versus *c* and *r*.

Where $\alpha \in (0,\ 1)$ is a constant, C is the number of clusters, $q^k_{V_i}$ is the scaled cluster indicator of P_i. S_i denotes the normalized transition matrix, i.e., $S_i(u,t) = P_i(u,t) \Big/ \sum_{l=1}^{K+1} P_i(u,l)$. $c^K_{V_i}$ is a vector including $K+1$ elements. Here, $\bar{q}^k_i{=}c^k_{V_i}[K+1]$ is the estimate of how likely it is that node i belongs to the k-th cluster. The goal is to maximize the concordance between $\bar{q}^k_i$ and q^k_i. Let $\beta_i \in R^{K+1}$ be the *i-th* row of the matrix $(1-\alpha)(I-\alpha S_i)^{-1}$, representing label propagation at its terminal state. We set $\bar{q}^k_i = \beta_i q^k_{V_i}$. Thus, $\bar{q}^k_i$ can be approximated to:

$$\bar{q}^k_i \approx \frac{\beta_i[1:K]\, q^k_{N(i)}}{1-\beta_i[K+1]}. \tag{9}$$

Where $\beta_i[1:K]$ denotes the first K elements of β_i and $\beta_i[K+1]$denotes the $(K+1)$-th element in β_i.

Next, we used matrix B to represent the linear relationship: $\bar{q}^k \approx Bq^k, \quad k = 1, 2, \ldots, C$:

$$B_{ij} = \begin{cases} \frac{\beta_i[j]}{1-\beta_i[K+1]} & \text{if } x_j \in N(i) \text{ and } x_j \text{ is the } j\text{—th element in } N(i)\,; \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

To minimize the difference between $\bar{q}^k$ and q^k, an objective function can be defined as follows:

$$\sum_{i=1}^{n}\sum_{k=1}^{C}\left(\bar{q}^k_i - q^k_i\right)^2 = \sum_{k=1}^{C}|\bar{q}^k - q^k|^2 \approx \sum_{k=1}^{C}|q^k - Bq^k|^2$$
$$= Tr\left(Q^T(I-B)^T(I-B)\,Q\right). \tag{11}$$

Here, $Tr(\bullet)$ denotes the trace of a matrix. Setting $Vir = (I-B)^T(I-B)$, we thus obtain the Vicus matrix. In this study, we propose to exploit the Vicus matrix as a graph regularization term to enhance the prediction performance of ILMF.

Note that each item in the Vicus matrix obtained from Eq. 11 represents the probability of vertex *i* having the same label as vertex *j*. Encoding the local neighborhood of each vertex in this way does not only preserve the geometric attributes of the Laplacian matrix but also improves the quality of clustering (Nelson et al., 2019). Wang B. et al. (2017) indicated the Vicus-based spectral clustering approach outperformed Laplacian-based methods on many biological tasks, such as single-cell RNA data clustering, recognition of rare cell populations, the ranking of genes related to cancer subtypes and so on. Therefore, in this manuscript, we use Vicus spectral matrix to model fine-grained connections between drugs and metabolites.

Vicus Regularization Based Inductive Logistic Matrix Factorization

The final drug-metabolite association prediction model can be constructed by considering the existing drug-metabolite links and the local geometrical structure of drugs and metabolites. By introducing Vicus regularization into Eq. 7, the proposed ILMF method is formulated as follows:

$$\begin{aligned}\min_{U,V}\sum_{i=1}^{n}\sum_{m=1}^{m}\ &\left\{\left(1+cy_{ij}-y_{ij}\right)\log\left[1+\exp\left(F_{i.}^{d}UV'F_{.j}^{m}\right)\right]\right.\\ &\left.-cy_{ij}\left(F_{i.}^{d}UV'F_{.j}^{m}\right)\right\}+\frac{\lambda}{2}\left(|U|_F^2+|V|_F^2\right)+\\ &\frac{\phi}{2}\left[tr\left(\left(F^mU\right)'Vir^m\left(F^mU\right)\right)\right.\\ &\left.+tr\left(\left(F^dV\right)'Vir^d\left(F^dV\right)\right)\right].\end{aligned}\tag{12}$$

Where ϕ is a graph regularization parameter. Vir^m is the Vicus matrix of metabolites, and Vir^d is the Vicus matrix of drugs. Note that, in this study, we exploit the cosine similarity of the low-dimensional feature matrix of metabolites F^m (or drugs_{F^d}) to compute the Vicus matrix Vir^m or Vir^d, respectively.

The optimization problem in Eq. 12 can be solved by an alternating gradient ascent scheme. In particular, we adopt the AdaGrad algorithm (Duchi et al., 2011) to update *U* and *V*. Further details can be found in the study by Liu et al. (2016). Once the projection matrices *U* and *V* have been obtained, the association probability of any drug-metabolite pair can be predicted by Eq. 5. However, for many unobserved interaction pairs, the learned latent representation of drugs and metabolites may not be accurate since they are only based on unknown drug-metabolite pairs.

To address this problem, we adopted the practices outlined in other literature (Ma et al., 2020a). Let $\text{N}_d^+ = \{m_i \left| \sum_j y_{ij} > 0\right.\}$ and $N_m^+ = \{m_j \left|\sum_i y_{ij} > 0\right.\}$ denote the sets of observed drugs and metabolites, respectively. $N_d^+\left(d_i\right)$ denotes the set of *K* nearest neighbors of d_i in N_d^+. Similarly, $N_m^+\left(m_j\right)$ denotes the set of *K* nearest neighbors of m_j in N_m^+. We can replace the latent vector representation of a drug or metabolite with the representations of its neighbors. Then, for each drug d_i, the revised $\bar{w}_i$ is defined as:

$$\bar{w}_i = \left\{\begin{array}{ll} w_i, & \text{if} \quad d_i \in N_d^+ \\ \frac{1}{Q_i^d}\sum_{l=1}^{K}\mu_l^d w_l, & \text{if} \quad d_i \notin N_d^+ \end{array}\right. . \tag{13}$$

Where $Q_i^d = \sum_{l=1}^{K}\alpha^{l-1}S^d\left(d_i, d_l d_l\right)$ is a normalized term, $S^d = \text{cosine}\left(F^d, F^d\right)$ denotes the consensus drug–drug similarity matrix derived from multiple similarity networks. d_l indicates the *l*-th neighbor in $N_d^+\left(d_i\right)$ sorted in descending order according to the similarity with d_i. $\alpha \in [0, 1]$ is a decay factor, and $\mu_l^d = \alpha^{l-1}S^d(d_i, d_l)$ is a weight factor. Similarly, we can also obtain the optimal latent representation $\bar{m}_j$ for each metabolite m_j :

$$\bar{h}_j = \left\{\begin{array}{ll} h_j, & \text{if } m_j \in N_m^+ \\ \frac{1}{Q_i^m}\sum_{l=1}^{K}\mu_l^m h_l, & \text{if } m_j \notin N_m^+ \end{array}\right. . \tag{14}$$

Where $Q_j^m = \sum_{l=1}^{K}\alpha^{l-1}S^m\left(m_j, m_l\right)$, $S^m = \text{cosine}\left(F^m, F^m\right)$ indicates the consensus metabolite–metabolite similarity matrix. m_l is the *l*-th neighbor in $N_m^+\left(m_j\right)$, which is sorted in descending order according to similarity with m_j. $\mu_l^m = \alpha^{l-1}S^m\left(m_j, m_l\right)$ is a weight factor.

Finally, the interaction probability of a drug-metabolite pair is redefined as follows:

$$\bar{p}_{ij} = \frac{\exp\left(\bar{w}_i\bar{h'}_j\right)}{1+\exp\left(\bar{w}_i\bar{h'}_j\right)}. \tag{15}$$

To demonstrate the flowchart of ILMF, the pseudocode of ILMF is given in **Table 1**.

RESULTS AND DISCUSSION

Experimental Settings

Following the previous studies (Zheng et al., 2013; Ding et al., 2014; Liu et al., 2016; Zhang et al., 2018a,b, 2019, 2020; Ma et al., 2020a), the performance of various association prediction methods can be evaluated by performing fivefold cross-validation (CV). For each method, we perform fivefold CV five times. Then, we calculate the area under the receiver operating characteristic curve (AUC), the area under the precision-recall curve (AUPR) scores in each repetition of CV, and the final AUC and AUPR scores are obtained by calculating the average over the five repetitions.

The object of this study is to predict the latent drug-metabolite associations. For the known drug-metabolite interaction matrix $Y \in R^{n\times m}$ with *n* drugs and *m* metabolites, we conduct CV on randomly selected drug-metabolite pairs. Specifically, we randomly divide the observed and unobserved interaction pairs into five equal parts. Then, in each round, one is used as test data, the remaining entries in *Y* are used for training. Thus, each of the five test datasets (or training data) includes the same number of observed and unobserved interaction pairs.

TABLE 3 | Top 20 novel associations predicted by ILMF on the "DrugMetaboliteAtlas" dataset.

Rank	Drug category	Metabolite	Score	Evidence (ATC/drug name)
1	C_HMG CoA reductase inhibitors-hydrophilic statin	TotPG	0.9915	C10AA03 (pravastatin)
2	M_Preparations inhibiting uric acid production	L.VLDL.FC	0.9891	M04AA01 (allopurinol)
3	M_Preparations inhibiting uric acid production	L.VLDL.P	0.9881	Unconfirmed
4	N_Benzodiazepine derivatives	UnsatDeg	0.9755	N03AE01 (clonazepam)
5	C_Angiotensin II antagonists-plain	XS.VLDL.FC	0.9687	Unconfirmed
6	C_Low-ceiling diuretics	XL.HDL.FC	0.9625	C03AA04 (chlorothiazide)
7	C_Low-ceiling diuretics	L.HDL.P	0.9588	C03AA03 (hydrochlorothiazide)
8	C_Low-ceiling diuretics	L.HDL.PL	0.9553	Unconfirmed
9	A_Insulins and analogs-fast-acting	FALen	0.9525	A10AB019 (insulin)
10	C_Low-ceiling diuretics	HDL.C	0.9493	Unconfirmed
11	B_Carbasalate calcium	ApoB	0.9419	Unconfirmed
12	C_Low-ceiling diuretics	HDL2.C	0.9346	Unconfirmed
13	C_Low-ceiling diuretics	UnsatDeg	0.9334	C03AA03 (hydrochlorothiazide)
14	C_Digoxin	S.VLDL.PL	0.9247	Unconfirmed
15	C_ACE inhibitors-plain	M.HDL.C	0.9240	C09AA01 (captopril)
16	C_HMG CoA reductase inhibitors-hydrophilic statin	S.HDL.CE	0.9219	C10AA03 (pravastatin)
17	C_Angiotensin II antagonists-plain	L.HDL.TG	0.9212	Unconfirmed
18	C_Fibrates	VLDL.D	0.9192	Unconfirmed
19	C_Angiotensin II antagonists-plain	PUFA	0.9188	C09CA01-08
20	M_Preparations inhibiting uric acid production	XL.VLDL.PL	0.9158	Unconfirmed

TotPG, total phosphoglycerides; L.VLDL.P, concentration of large VLDL particles; L.VLDL.FC, free cholesterol in very large VLDL; UnsatDeg, estimated degree of unsaturation; XS.VLDL.FC, free cholesterol in very small VLDL; XL.HDL.FC, free cholesterol in very large HDL; L.HDL.P, concentration of large HDL particles; L.HDL.PL, phospholipids in large HDL; FALen, estimated description of fatty acid chain length- not actual carbon number; HDL.C, total cholesterol in HDL; ApoB, apolipoprotein B; HDL2.C, total cholesterol in HDL2; S.VLDL.PL, phospholipids in small VLDL; M.HDL.C, total cholesterol in medium HDL; S.HDL.CE, cholesterol esters in small HDL; L.HDL.TG, triglycerides in large HDL; VLDL.D, mean diameter for VLDL particles; PUFA, polyunsaturated fatty acids; XL.VLDL.PL, phospholipids in very large VLDL; VLDL, very-low-density lipoprotein; HDL, high-density lipoprotein.

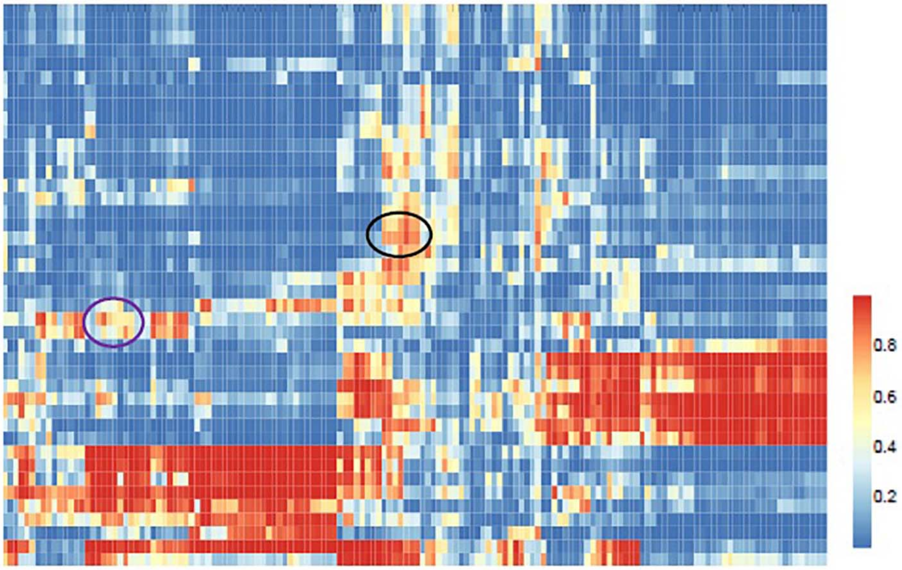


FIGURE 4 | Global view of the predicted drug-metabolite associations. Hierarchical clustering of the ILMF scores between 42 drugs and 150 metabolites. The color of each cell represents the ILMF score of a drug (row) and a metabolite (column), where red/blue indicates high/low ILMF scores.

Note that we do not consider the other two scenarios for CV experiments: random rows or columns selected for testing. It is mainly because the drug-metabolite association matrix is commonly sparse, and the drug–drug or metabolite–metabolite similarity information from external sources cannot provide enough aid for prediction.

Evaluation Metrics and Competing Approaches

In this study, the AUC, AUPR, and F1 value are used as the evaluation metrics. These metrics have been widely used in various association prediction tasks. To demonstrate the effectiveness and efficiency of our proposed ILMF algorithm in predicting drug-metabolite interaction, we compare the proposed ILMF method with the following several state-of-the-art approaches, namely, DTInet (Luo et al., 2017), IMCMDA (Chen et al., 2018) and GRNMF (Xiao et al., 2018). These approaches were originally designed for DTI prediction or miRNA-disease association prediction. Furthermore, we can obtain a variant of ILMF, which learns U and V with the consensus similarity matrices of drugs and metabolites instead of their Vicus matrices. Here, we denote this variant as ILMF^{-}, which has a similar objective function to MNLMF (Ma et al., 2020a) and NRLMF (Liu et al., 2016).

For all the compared methods above, their performance is reported with best-tuned parameters.

Experimental Results

In this subsection, we conduct extensive experiments on the “DrugMetaboliteAtlas” dataset. **Table 2** shows the performance of various algorithms in terms of AUC, AUPR, and F1. In **Table 2**, the highest score in each column is shown in bold typeface.

As shown in **Table 2**, ILMF achieves the best performance in terms of AUC, AUPR, and F1 on the “DrugMetaboliteAtlas” dataset. Specifically, compared with the second-best GRNMF algorithm, the performance of ILMF increases by 1.40, 7.80, and 4.94% in terms of AUC, AUPR, and F1, respectively. Additionally, the prediction performance of DTInet and IMCMDA is not satisfactory. We can observe from **Table 2** that ILMF outperforms IMCMDA 18.82, 72.45, and 39.29% in AUC, AUPR, and F1, respectively. One possible reason is that IMCMDA does not take advantage of the local geometrical structure that resided within the original data. For GRNMF, it does not consider the important level parameter c, for simplicity, it views the known drug-metabolite pairs and the unobserved drug-metabolite pairs as equally important in predicting the latent associations between drugs and metabolites.

By comparing ILMF and ILMF^{-}, we can also further verify the benefits of using the Vicus matrices of drugs and metabolites, indicating that exploiting the local structure information of drugs and metabolites could improve the performance for drug-metabolite association prediction.

Parameter Analysis

There are several parameters in ILMF that need to be tuned: the important level parameter c, the dimensionality k_1, k_2 and r of projection matrices W and H, the regularization parameters λ and ϕ. For simplicity, we set $k_1 = 12$ and $k_2 = 45$ empirically. We adopted a grid search strategy to select the optimal combination from fixed ranges of λ and ϕ. In this study, we let λ and ϕ vary in the range $\{2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}\}$, r varies in the range $\{5, 6, 7, 8, 9, 10, 11, 12\}$ and c varies in the range $\{2, 3, 4, 5, 6, 7, 8\}$. We then conducted fivefold CV to evaluate the performance of ILMF under the combination of different parameters.

To demonstrate how λ and ϕ affect the performance of the proposed ILMF, we fix other parameters and change the values of λ and ϕ, respectively. The AUC and AUPR scores are shown in **Figures 2A,B** with respect to different combinations of λ and ϕ .

λ and ϕ are the parameters controlling the influence of feature regularization and graph regularization. As **Figure 2** shows, when we fix the values of λ and increase the values of ϕ, the AUC scores increase initially and decrease after achieving the highest performance. These results demonstrate the advantages of introducing two kinds of regularization terms.

In this study, we also conducted extensive experiments to demonstrate how c and r affect the performance of ILMF. We changed the values of c and r in the corresponding ranges with other parameters fixed. The AUC and AUPR scores are shown in **Figures 3A,B** with respect to different combinations of c and r. We can observe from **Figure 3** that for a fixed value of c, the AUC scores increase as the values of r increase. However, when we fix the values of r and increase the values of c, the AUC scores decrease. Similar properties can be seen in terms of AUPR. This illustrates the importance and necessity of introducing levels of importance, which are assigned to the observed drug-metabolite interaction pairs.

PREDICTING NOVEL DRUG-METABOLITE ASSOCIATIONS

In this section, we evaluate the prediction ability of ILMF in identifying novel drug-metabolite associations. In our experiments, the entire dataset is used to train the ILMF model, and the optimal parameters are used to make a prediction. The unknown drug-metabolite interaction pairs are ranked based on the predicted association scores.

Table 3 shows the top 20 novel associations predicted by ILMF on the “DrugMetaboliteAtlas” dataset. In this table, the fourth column shows the predicted interaction probabilities of novel drug-metabolite pairs. For each pair, we retrieval the possible interaction from HMDB, DrugBank and other databases that may contain it, and list the corresponding ATC/drug names in the last column of **Table 3**. Since only a few databases include drug-metabolite association information, the fraction of new drug-metabolite interactions correctly predicted by ILMF may increase in the future. These promising results, which indicate that ILMF can successfully identify many novel associations, demonstrates that it is effective in predicting latent drug-metabolite associations from a sparse binary matrix.

Note that the proposed ILMF is also effective when a new drug (or metabolite) without any known related metabolites (or drugs) is given. Once we have obtained the low-dimensional matrix representation $F^{d}_{new(i)}$ of a new drug or $F^{d}_{new(j)}$ of a metabolite, the interaction scores with known drugs or metabolites can be calculated by Eq. 15.

We further apply ILMF to detect the relationships between drugs and metabolites from a global view. ILMF is used to infer the metabolic potential of 42 drugs and chart the

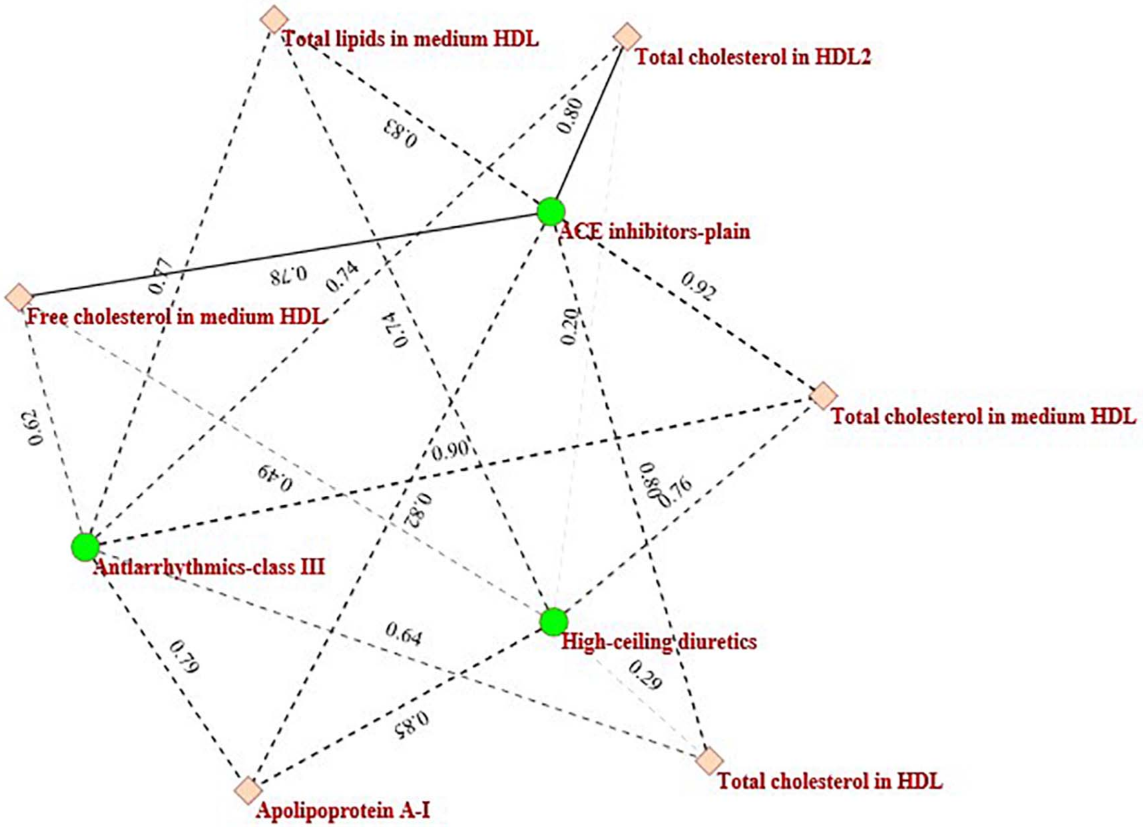


FIGURE 5 | The sub-network consists of three drugs and six metabolites.

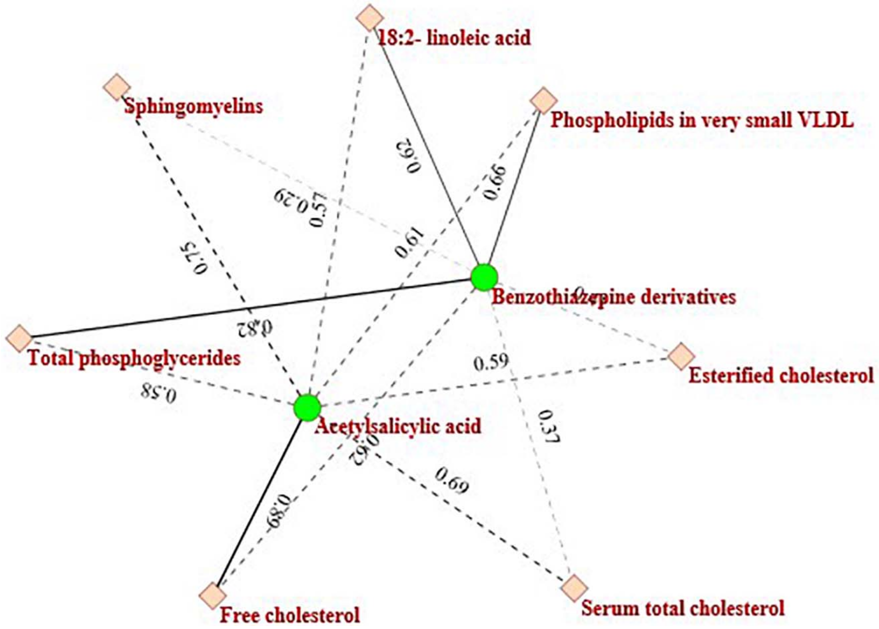


FIGURE 6 | The sub-network consists of two drugs and seven metabolites.

metabolic landscape of common drugs. First, we obtain a score matrix by applying ILMF on the whole "DrugMetaboliteAtlas" dataset. Then, hierarchical clustering is performed to explore the unknown relationships between drugs and metabolites (**Figure 4**). The scores indicate the interaction relationships between drugs and metabolites based on metabolic mechanisms. Therefore, the drugs and metabolites that are grouped may share metabolic overlaps in terms of pathways or microbial metabolites association profiles.

In **Figure 4** the black circled region shows a module that consists of three categories of drugs (*Antiarrhythmics-class III, ACE inhibitors-plain*, and *High-ceiling diuretics*) and six kinds of metabolites (*Total cholesterol in HDL2, Total cholesterol in HDL, Free cholesterol in medium HDL, Total cholesterol in medium HDL, Total lipids in medium HDL, and Apolipoprotein A-I*). These drugs and metabolites, which have no associations in the original drug-metabolite association matrix are identified by the proposed ILMF. The relationships between these drugs and metabolites have been reported in some literature. **Figure 5** shows the connectivity of this module by extracting the corresponding rows and columns from the predicted drug-metabolite scoring matrix. The green circle denotes the three drugs mentioned above. The pink diamond denotes six metabolites. Solid lines indicate the true associations between drugs and metabolites. Dot lines indicate the predicted associations by ILMF. The values on the lines are the predicted scores. The bigger the score, the more trustworthy the predicted drug-metabolite interaction pair. This setting is also applied to **Figure 6**.

As shown in **Figure 5**, *Total cholesterol in medium HDL* is highly related to *Antiarrhythmics-class III* and *ACE inhibitors-plain* and the predicted interaction scores between them are 0.90 and 0.92, respectively. This is consistent with the fact that high *Total cholesterol* level usually leads to other complications, including diabetes, hyperlipidemia, hypertension, hypothyroidism, choledochus obstruction, coronary heart disease, atherosclerosis, and so on (Nelson, 2013). Miyazaki et al. (1999) also reported that ACE activity was significantly increased in the aorta of cholesterol-fed monkeys.

Another example is the purple circled region, which contains two kinds of drugs (Antithrombotic agents-Acetylsalicylic acid: B01AC06 and Benzothiazepine derivatives: C08DB01) and seven metabolites (Sphingomyelins, Serum total cholesterol, Total phosphoglycerides, Esterified cholesterol, Free cholesterol, 18:2-linoleic acid, and Phospholipids in very small VLDL). The drugs and metabolites in this module are also clinically relevant. **Figure 6** describes the heterogeneous interaction network of this module. As **Figure 6** indicates, Acetylsalicylic acid is related to Sphingomyelins (interaction probability is 0.7531). This finding is also consistent with another previous report by Suwalsky et al. (2013).

There have also been other biologically meaningful modules detected by ILMF. In short, the two examples mentioned above show the potential of the proposed ILMF algorithm in identifying the unknown associations between drugs and metabolites, which further demonstrates its effectiveness and efficiency.

CONCLUSION

In this article, we propose a novel drug-metabolite association prediction method, named ILMF. ILMF could not only combine multiple-source drug–drug interaction, metabolite–metabolite interaction, and drug-metabolite association information into this framework but also take full advantage of the local geometrical structure inherent in the original data to improve prediction performance. In addition, we also exploited inductive matrix completion to guide the learning of projection matrices *U*, *V* based on the low-dimensional feature matrix of drugs (or metabolites) obtained from external data sources. The experimental results for the "DrugMetaboliteAtlas" dataset demonstrate the effectiveness of the proposed ILMF in predicting potential drug-metabolite associations. Moreover, in the last section of this study, we examine case studies on predicting novel drug-metabolite associations, the results of which may provide some valuable clues to biologists or clinicians.

Despite these promising findings, there are still some limitations to this proposed ILMF model. While fusing multiple types of biological data, the chemical structure information of drugs or metabolites is missing due to the fact that the initial "DrugMetaboliteAtlas" dataset only contains vague categories, particularly for metabolites. The low-dimension feature representation learning algorithm (clusDCA) is replaceable. More effective graph representation learning frameworks, such as graph convolution network (GCN), are expected to be combined with the ILMF framework to more accurately predict drug-metabolite associations. Lastly, the predicted drug-metabolite interactions need to be further validated in practice.

In the future, we will focus on developing new methods to explore the complex relationships between drugs and microbes, including the influence of microbes on drug activity or toxicity and so on.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/chonghua-1983/ILMF.

AUTHOR CONTRIBUTIONS

YYM wrote the manuscript and developed the algorithms. YYM and LL developed the concept for the structure and content of the manuscript. QC wrote the code used in the manuscript. YJM critically revised the final manuscript. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61532008), Key Research Projects of Henan Higher Education Institutions (No. 20B520002), and The Key Technology R&D Program of Henan Province (No. 202102310561).

REFERENCES

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Cho, H., Berger, B., and Peng, J. (2015). *Diffusion component analysis: unraveling functional topology in biological networks. in: International Conference on Research in Computational Molecular Biology*. Berlin: Springer, 62–64.

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings Bioinform.* 15, 734–747. doi: 10.1093/bib/bbt056

Doestzada, M., Vila, A. V., Zhernakova, A., Koonen, D. P., Weersma, R. K., Touw, D. J., et al. (2018). Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein cell* 9, 432–445. doi: 10.1007/s13238-018-0547-2

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159.

Fuller, A. (1937). Is p-aminobenzenesulphonamide the active agent in prontosil therapy? *Lancet* 229, 194–198. doi: 10.1016/s0140-6736(00)97447-6

Haiser, H. J., Gootenberg, D. B., Chatman, K., Sirasani, G., Balskus, E. P., and Turnbaugh, P. J. (2013). Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. *Science* 341, 295–298. doi: 10.1126/science.1235872

He, B.-S., Peng, L.-H., and Li, Z. (2018). Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front. Microbiol.* 9:2560. doi: 10.3389/fmicb.2018.02560

Hu, Y., Koren, Y., and Volinsky, C. (2008). *Collaborative filtering for implicit feedback datasets. in: 2008 Eighth IEEE International Conference on Data Mining*. Pisa, Italy: IEEE, 263–272.

Jain, P., and Dhillon, I. S. (2013). *Provable inductive matrix completion. arXiv preprint arXiv:1306.0626.* **.

Johnson, C. C. (2014). Logistic matrix factorization for implicit feedback data. *Adv. Neural Inform. Proc. Syst.* 27, 1–9.

Kurilshikov, A., Van Den Munckhof, I. C., Chen, L., Bonder, M. J., Schraa, K., Rutten, J. H., et al. (2019). Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk: a cross-sectional study. *Circ. Res.* 124, 1808–1820. doi: 10.1161/circresaha.118.314642

Liu, J., Lahousse, L., Nivard, M. G., Bot, M., Chen, L., Van Klinken, J. B., et al. (2020). Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas. *Nat. Med.* 26, 110–117.

Liu, Y., Wei, W., Sun, A., and Miao, C. (2014). *Exploiting geographical neighborhood characteristics for location recommendation. in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*). New York: ACM, 739–748.

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* 12:e1004760. doi: 10.1371/journal.pcbi.1004760

Long, Y., and Luo, J. (2020). Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. *IEEE J. Biomed. Health Inform.* 25, 266–275. doi: 10.1109/jbhi.2020.2998906

Long, Y., Wu, M., Kwoh, C. K., Luo, J., and Li, X. (2020).). Predicting human microbe-drug associations via graph convolutional network with conditional random field. *Bioinformatics* 36, 4918–4927. doi: 10.1093/bioinformatics/btaa598

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8, 1–13.

Ma, Y., He, T., and Jiang, X. (2020a). Multi-network logistic matrix factorization for metabolite–disease interaction prediction. *FEBS Lett.* 594, 1675–1684. doi: 10.1002/1873-3468.13782

Ma, Y., Liu, G., Ma, Y., and Chen, Q. (2020b). Integrative analysis for identifying co-modules of microbe-disease data by matrix tri-factorization with phylogenetic information. *Front. Genet.* 11:83. doi: 10.3389/fgene.2020.00083

Miyazaki, M., Sakonjo, H., and Takai, S. (1999). Anti-atherosclerotic effects of an angiotensin converting enzyme inhibitor and an angiotensin II antagonist in cynomolgus monkeys fed a high-cholesterol diet. *Br. J. Pharmacol.* 128, 523–529. doi: 10.1038/sj.bjp.0702833

Natarajan, N., and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* 30, i60–i68.

Nelson, R. H. (2013). Hyperlipidemia as a risk factor for cardiovascular disease. *Primary Care* 40, 195–211. doi: 10.1016/j.pop.2012.11.003

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., and Sharan, R. (2019). To embed or not: network embedding as a paradigm in computational biology. *Front. Genet.* 10:381. doi: 10.3389/fgene.2019.00381

Shang, D., Li, C., Yao, Q., Yang, H., Xu, Y., Han, J., et al. (2014). Prioritizing candidate disease metabolites based on global functional relationships between metabolites in the context of metabolic pathways. *PloS One* 9:e104934. doi: 10.1371/journal.pone.0104934

Suwalsky, M., Belmar, J., Villena, F., Gallardo, M. J., Jemiola-Rzeminska, M., and Strzalka, K. (2013). Acetylsalicylic acid (aspirin) and salicylic acid interaction with the human erythrocyte membrane bilayer induce in vitro changes in the morphology of erythrocytes. *Arch. Biochem. Biophys.* 539, 9–19. doi: 10.1016/j.abb.2013.09.006

Wang, B., Huang, L., Zhu, Y., Kundaje, A., Batzoglou, S., and Goldenberg, A. (2017). Vicus: exploiting local structures to improve network-based analysis of biological data. *PLoS comput. Biol.* 13:e1005621. doi: 10.1371/journal.pcbi.1005621

Wang, F., Huang, Z.-A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* 7, 1–11.

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31, i357–i364.

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. J. B. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., et al. (2015). Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Sci. Rep.* 5:17201.

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi: 10.1039/c9sc04336e

Zhang, W., Chen, Y., Li, D., and Yue, X. (2018a). Manifold regularized matrix factorization for drug-drug interaction prediction. *J. Biomed. Inform.* 88, 90–97. doi: 10.1016/j.jbi.2018.11.005

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018b). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinform.* 19:233. doi: 10.1186/s12859-018-2220-4

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inform. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, Z.-C., Zhang, X.-F., Wu, M., Ou-Yang, L., Zhao, X.-M., and Li, X.-L. (2020). A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics* 36, 3474–3481. doi: 10.1093/bioinformatics/btaa157

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). *Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*). New York: ACM, 1025–1033.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. *Adv. Neural Inform. Proc. Syst.* 16, 321–328.

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.

frontiers in Microbiology

METHODS
published: 05 May 2021
doi: 10.3389/fmicb.2021.640693

HCK and ABAA: A Newly Designed Pipeline to Improve Fungi Metabarcoding Analysis

*Kodjovi D. Mlaga[1], Alban Mathieu[1,2], Charles Joly Beauparlant[1,2], Alban Ott[3], Ahmad Khodr[3], Olivier Perin[3] and Arnaud Droit[1,2]**

[1] Department of Molecular Medicine, Laval University, Quebec, QC, Canada, [2] Centre de Recherche du CHU de Québec, Quebec, QC, Canada, [3] Research and Innovation, L'Oreal, Paris, France

Introduction: The fungi ITS sequence length dissimilarity, non-specific amplicons, including chimaera formed during Polymerase Chain Reaction (PCR), added to sequencing errors, create bias during similarity clustering and abundance estimation in the downstream analysis. To overcome these challenges, we present a novel approach, Hierarchical Clustering with Kraken (HCK), to classify ITS1 amplicons and Abundance-Base Alternative Approach (ABAA) pipeline to detect and filter non-specific amplicons in fungi metabarcoding sequencing datasets.

Materials and Methods: We compared the performances of both pipelines against QIIME, KRAKEN, and DADA2 using publicly available fungi ITS mock community datasets and using BLASTn as a reference. We calculated the Precision, Recall, F-score using the True-Positive, False-positive, and False-negative estimation. Alpha diversity (Chao1 and Shannon metrics) was also used to evaluate the diversity estimation of our method.

Results: The analysis shows that ABAA reduced the number of false-positive with all metabarcoding methods tested, and HCK increases precision and recall. HCK, coupled with ABAA, improves the F-score and bring alpha diversity metric value close to that of the BLASTn alpha diversity values when compared to QIIME, KRAKEN, and DADA2.

Conclusion: The developed HCK-ABAA approach allows better identification of the fungi community structures while avoiding use of a reference database for non-specific amplicons filtration. It results in a more robust and stable methodology over time. The software can be downloaded on the following link: https://bitbucket.org/GottySG36/hck/src/master/.

Keywords: ABAA, benchmarking, F-score, fungi, HCK, hierarchical clustering, ITS amplicons

OPEN ACCESS

Edited by:
Liang (Leon) Wang,
Xuzhou Medical University, China

Reviewed by:
Oliver Hawlitschek,
University of Hamburg, Germany
Hideyuki Doi,
University of Hyogo, Japan

***Correspondence:**
Arnaud Droit
Arnaud.Droit@crchudequebec.ulaval.ca

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

Received: *11 December 2020*
Accepted: *08 April 2021*
Published: *05 May 2021*

Citation:
Mlaga KD, Mathieu A, Beauparlant CJ, Ott A, Khodr A, Perin O and Droit A (2021) HCK and ABAA: A Newly Designed Pipeline to Improve Fungi Metabarcoding Analysis. Front. Microbiol. 12:640693. doi: 10.3389/fmicb.2021.640693

INTRODUCTION

The mycobiome concept was first introduced in 2010 to designate the fungal community of the human oral cavity (Tang et al., 2015) before being extended to other micro-environments. Three genomic markers are widely used to identify fungal species in a microbial environment: 18S ribosomal gene (Wu et al., 2015), 28S ribosomal gene (Ninet et al., 2003), and the Internal

Transcribed Spacers (ITS) (Martin and Rygiewicz, 2005; Bellemain et al., 2010). The most commonly used is the ITS amplicon (Fujita et al., 2001) which targets two loci: ITS1, located between the 18S and 5.8S genes, and ITS2, between 5.8S and 28S (Bellemain et al., 2010). ITS1 has been demonstrated to yield the best performance (Bazzicalupo et al., 2013; Wang et al., 2015). Several packages have been developed to automate the process, and most of them are OTU (Operational Taxonomic Unit) sequence similarity-based pipeline (Schloss et al., 2009; Gweon et al., 2015; Rognes et al., 2016; Mysara et al., 2017; Bolyen et al., 2018). To date, the research communities are gradually moving to the new concept of ASVs (Amplicons sequence Variants) or Exact Sequences Variants (ESVs) (Callahan et al., 2017). With these pipelines, the taxonomy delineates based on the single nucleotides' variant of amplicons, assuming that amplicons sequences have a similar length which is not the case with fungi ITS sequences. To date, several pipelines have been developed to classify fungal species using ITS sequencing. These include Plutof (Abarenkov et al., 2010b), Clotu (Kumar et al., 2011), PIPITS (Gweon et al., 2015), CloVR-ITS (White et al., 2013), and BioMaS (Fosso et al., 2015) specially designed to analyse fungi ITS datasets, Kraken (Wood and Salzberg, 2014), Mothur (Schloss et al., 2009) Qiime (Caporaso et al., 2010; Bolyen et al., 2018), Vsearch (Rognes et al., 2016), and DADA2 (Callahan et al., 2016) among many others, to examine both bacterial 16S rRNA and fungal ITS amplicons.

The size of fungal ITS sequences is highly variable, and species can differ widely by the number of loci (Tang et al., 2015; Khodadadi et al., 2017). The sequence length dissimilarity creates bias during clustering and affects OTUs abundance estimation. Moreover, besides biologically valid amplicons, PCR generates many non-specific fragments resulting from elongation interruption or two or more incomplete amplicons joining (chimaeras) (Lahr and Katz, 2009; Edgar, 2016; Bjørnsgaard Aas et al., 2017). These non-specific amplicons are hybrid products between multiple parent sequences that can be falsely interpreted as existing or novel species, thus significantly affect the diversities, including the alpha and beta diversity metrics (Zajec et al., 2012). Hence, non-specific amplicons formed during amplification with two incomplete segments (*bimeras*) are generally at a lower proportion. However, chimaeras with more than two fragments (*multimers*) may form at comparable rates and account for a significant fraction in an amplified sample (Lahr and Katz, 2009). The most commonly used pipeline to detect chimaeras is UCHIME, composed of reference-based and *de novo* approaches (Edgar, 2016). The reference-based approach detects non-specific amplicons in a dataset by making a model from a concatenated pair of sub-sequences in a reference database. Chimaeras are detected if the query alignment sequence score of the model exceeds a threshold. UCHIME depends on a reference database, and ITS sequence size variation can be a significant source of false-positive detection, throwing away biologically valid sequences. DADA2 implements *isBimeraDenovo()* function that identifies exact bimeras or multimeras sequences. Child sequences that differ by a single mismatch from the chimeric model are flagged if the left parent and right parent are at least four nucleotides away from the child sequence (Callahan et al., 2016). The challenge is that databases are rarely updated, and the similarity search can be time-consuming, especially when databases are large. Computational resources are one of the critical limitations. Maintaining specific databases up to date is a real challenge, and a broad range of databases suffer from contamination and unannotated sequences. The available databases, such as UNITE, which is commonly used, presents 26% of entries that cannot be consistently assigned to a taxonomic family (Nilsson et al., 2008; Kõljalg et al., 2013). These tools are mainly developed for 16S/18S markers but widely applied to fungal ITS amplicons. Besides, these tools have been optimised using simulated datasets and not real datasets (Bjørnsgaard Aas et al., 2017).

To overcome the above limitations, we present a novel classification approach for ITS amplicon's taxonomy assignment. This approach consists of two steps: The amplicons Abundance-Base Alternative Approach (ABAA), a *de novo* method to filter non-specific amplicons from sequence datasets and a Hierarchical Clustering with Kraken (HCK) to classify ITS amplicons. We built HCK on a hierarchical clustering approach with multiple-step iterating runs. Each cluster's representative sequences are taxonomically assigned using *Kraken* with the exact alignment of k-mers using fungal ITS loci sequence database (ITSdb). In this study, we use comparative analysis approach to assess the performance of ABAA and HCK. We calculated the Precision, the Recall, and the F-score using the True-Positive, False-positive, and False-negative estimation. Alpha diversity (Chao1 and Shannon) was also used to evaluate the methods' diversity estimation. Chao1 is based on the concept that rare species allow inferring the number of missing species. As the Chao1 richness estimator gives more weight to the low abundance species while the Shannon index measures the richness and the evenness (Kim et al., 2017), making the Chao1 metric more sensitive to abundance estimation than Shannon's. Henceforth, to simplify the manuscript, chimaeras and non-specific amplicons will interchangeably be used to designate all non-specific amplicons, including chimaeras, incomplete amplicons and sequencing errors.

MATERIALS AND METHODS

The methodology in this study is organised in two parts. In the first part, we will describe ABAA and HCK workflow using publicly available ITS mock community datasets. We will then, in a second part, compare the performance of HCK-ABAA to that of QIIME, DADA2 and KRAKEN using BLASTn search abundance estimation as a reference.

Fungi ITS Mock Communities' Datasets

We downloaded Biological mock community datasets of three different projects from the SRA NCBI database. The three projects were conducted using the Illumina Miseq sequencing technology. The first project, available under accession number *PRJNA516455* (McTaggart et al., 2019), contains six different samples (*SRR8473974, SRR8473977, SRR8473978, SRR8473979, SRR8473980, SRR8473984*), which were prepared from subsets of

53 species of fungi with an emphasis on human lung pathogens. The second project, available under accession number *SRP132544* (Hoggard et al., 2018), contains three samples (*SRR6702280, SRR6702281, SRR6702283),* including specific fungal species from different human body location or organs (lung, oral cavity, gastrointestinal tract, and skin). The third project, available under accession number *PRJNA382746*, contains two samples (*SRR5439721, SRR5439722)* that include 16 species of fungi. Overall, the mock communities contain 36 fungi genera which are: *Alternaria, Apophysomyces, Aspergillus, Blastomyces, Candida, Cladosporium, Clavispora, Coccidioides, Cryptococcus, Cunninghamella, Exophiala, Fusarium, Histoplasma, Lichtheimia, Malassezia, Meyerozyma, Mucor, Paecilomyces, Penicillium, Phanerochaete, Pichia, Purpureocillium, Rasamsonia, Rhizopus, Saccharomyces, Sarocladium, Scedosporium, Schizosaccharomyces, Sporidiobolus, Sporothrix, Talaromyces, Trichoderma, Trichosporon, Wickerhamomyces, Sclerotina, Rhyzomucor, Trichophyton* detailed in **Table 1**.

Fungi ITS Analysis Workflow With HCK-ABAA

Data Pre-processing and Quality Check

The sequence reads are trimmed with paired-end mode using *Trimmomatic* (Bolger et al., 2014) to remove residual adapters. The default parameters are used, including “phred33” to encode the quality part of the Fastq file to base 33, the low-quality bases from the sequence beginning and the end is set to 3 bases, respectively. The sliding window size was set to 4 with a minimum length of 50 bases. The paired reads generated from the trimming are then joined into contigs to produce the final fasta file using *Pandaseq* (Masella et al., 2012) with default parameters. Sequences with ambiguous bases are removed.

Non-specific Amplicons Filtering: ABAA

We empirically consider that amplicons with length-frequency below the standard deviation overall distribution to originate from non-specific amplification. Technically, after determining the amplicons’ length distribution and their frequency within each sample, an amplicon is considered to be non-specific if its length-frequency is below a certain threshold. This threshold corresponds to the standard deviation of the frequency of the amplicon lengths. ABAA filtering corresponds to step 1 of the whole pipeline.

Hierarchical Clustering With Kraken Assignment (HCK)

Amplicons Hierarchical Clustering

Amplicon hierarchical clustering corresponds to step 2 of the whole pipeline. HCK clusters amplicons sequences using multiple-step iterated runs of sequence alignments with a neighbour-joining algorithm implemented in CD-HIT version 4.5.4 (Fu et al., 2012). A segment sliding window in this context or “word” is defined as the consecutive position of a certain number of nucleotides in a sequence fragment. We implemented three iterative runs in the clustering and set the sequence identities (c) to 0.99, 0.98, and 0.97, as well as the “*word*” size (n) to 10, 8, and 7 bps, respectively. It is possible to control the sequence length difference cut-off(s), the alignment coverage of the more extended sequence (aL), and the alignment coverage for the shorter sequence (aS). The most crucial parameter is the length difference cut-off(s) depending on the overall distribution of the amplicon’s size. It can be empirically estimated by dividing the average size by the size of the most extended amplicon. This value was set to 90% in the study. The iterated clusters generated are then merged into one single, no redundant cluster file and sorted by size to remove singleton amplicons. An intermediary step 3 is essential to retrieve representative sequences from each cluster and be classified using Kraken (**Figure 1A**).

ITS Loci RefSeq

We downloaded the fungal Internal Transcribed Spacer RNA (ITS) RefSeq Targeted Loci (ITSdb) containing 11,252 entries. We retrieved the corresponding taxonomy profile from the NCBI taxonomy database[1] and created a Qiime-compatible taxonomy file. Both files (fasta and taxonomy file) were sorted and cleaned to have similar entries, using the following utilities[2]. ITSdb was used to generate a kraken database following the procedure available at this web address: http://ccb.jhu.edu/software/kraken/MANUAL.html.

Taxonomical Classification

Each cluster’s representative sequences are classified using the Lowest Common Ancestor (LCA) algorithm with Kraken version 1 (Wood and Salzberg, 2014). The taxonomy assignment is then extended to other amplicons of the respective clusters for a complete classification. This step corresponds to step 4 of the HCK workflow. The command uses the sample metadata information to generate a BIOM file. The final stage, step 5, uses the BIOM file to estimate the diversity abundance and further metric calculation analysis (**Figure 1B**).

Benchmark Analysis and Performances Evaluation

BLASTn (Reference)

We determined the actual reference diversity and abundance with BLASTn sequence similarity search against the NCBI NT database. Consensus classification was determined for coverage ≥98%, identity ≥97%, and *e*-value ≤ 0.00001 with a maximum of 100 hits retained per entry. The BLASTn output was then filtered for the best hits successively by the *e*-value, coverage percentage, and identity percentage. The final consensual taxonomy classification for each amplicon is kept based on a minimum number of 80 identical taxid out of 100 for each query (80% of the total hits) to generate an abundance table following a procedure described by other authors (Blaalid et al., 2013; McTaggart et al., 2019).

Comparative Analysis

To evaluate the efficacy of the newly developed tools, we compared the absolute count diversity of HCK to Qiime v1.9

[1] ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Fungi/

[2] https://github.com/bakerccm/entrez_qiime

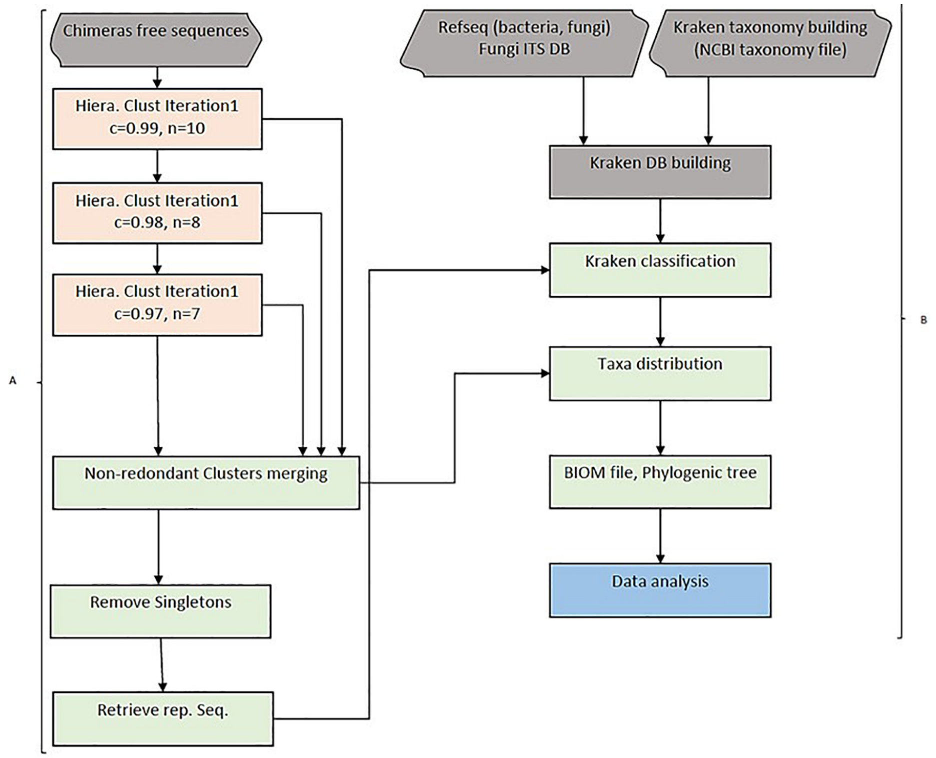

Chimeras free sequences
Hiera. Clust Iteration1 c=0.99, n=10
Hiera. Clust Iteration1 c=0.98, n=8
Hiera. Clust Iteration1 c=0.97, n=7
Non-redondant Clusters merging
Remove Singletons
Retrieve rep. Seq.
Refseq (bacteria, fungi) Fungi ITS DB
Kraken taxonomy building (NCBI taxonomy file)
Kraken DB building
Kraken classification
Taxa distribution
BIOM file, Phylogenic tree
Data analysis
A
B

FIGURE 1 | HCK workflow diagram. **(A)** Hierarchical clustering with three iterations. Chimaeras free sequences are results of pipeline step 1, including raw reads trimming, merging (forward and reverse reads) and ABAA filtering. Sequences are combined into one single fasta file and clustered using a hierarchical clustering approach in step 2. All clusters are then merged into one single non-redundant clusters and got rid of singletons sequences. HCK retrieves representative sequences from each cluster for amplicons' classification, the second part of the pipeline (step 3). **(B)** Classification uses Lowest Common Ancestor (LCA) taxonomical assignment implemented in Kraken to classify representative sequences and taxonomy reported to each cluster, and a final BIOM file can be generated for downstream analysis (Steps 4 and 5).

(Caporaso et al., 2010), Kraken (Wood and Salzberg, 2014), and DADA2 (version 1.8) (Callahan et al., 2016) with and without non-specific sequences/chimaera removal using BLASTn abundance estimation as reference. We test HCK, Kraken, Qiime with ITSdb, Qiime with UNITE (Abarenkov et al., 2010a) and ITSdb database. DADA2 is tested only with the native UNITE database. The performance of each method is determined by its ability to assign the suitable taxa to the right sequence and to be able to assign the maximum of good sequences using sensitivity (recall), the positive predictive value (precision), and the f-score metric calculation (**Figure 2**). We determined True positive (TP) as following: For x_i, the abundance estimated by the BLAST (reference) and y_i, the abundance estimated by the tested methods for given sample i, we determined true positives by $TP_i = min(x_i, y_i)$. The overestimated abundance classified by the tested method is considered false positive, and the underestimation differences are included in the false negatives. The false negatives (FN) are determined by the sum of counts of amplicon only detected by BLAST but are not correctly assigned by the assessed method. For Tr_i, the total abundance estimated by BLAST for given sample i, $FN_i = Tr_i - TP_i$. The false positive (FP) corresponds to the sum of counts of amplicons wrongly assigned by the tested method but not detected by BLAST or not included in the initial mock community composition. For Tm_i, the total number of amplicons classified for given taxa by the tested method, $FP_i = Tm_i - TP_i$. We determined the precision (P_i) and the recall (R_i) and calculated the F- score using the following formula. $P_i = TP_i/(TP_i+FP_i)$, $R_i = TP_i/(TP_i+FN_i)$, $F\text{-}score_i = 2^{*}P_{i*}R_i/(P_i + R_i)$ (Gardner et al., 2019). We also calculated alpha diversity using Shannon and chao1 indices to assess the association of chimaera removal methods and taxonomy classification in downstream diversity analysis. We compared it to the diversity of BLAST abundance estimation. We estimate the difference between the alpha diversity of the assessed methods and that of the BLASTn estimation. The lower the difference, the best is the method. All scripts and command lines are details in **Supplementary Material**: scripts_and_command_lines.

RESULTS

Fungi ITS Amplicons Length: A Vast Diversity Among Species

All 11 samples from the three projects were combined into one single dataset during the pre-processing treatment. The average read length is 200.9 bp (*SD* = 65.6), the maximum read size is

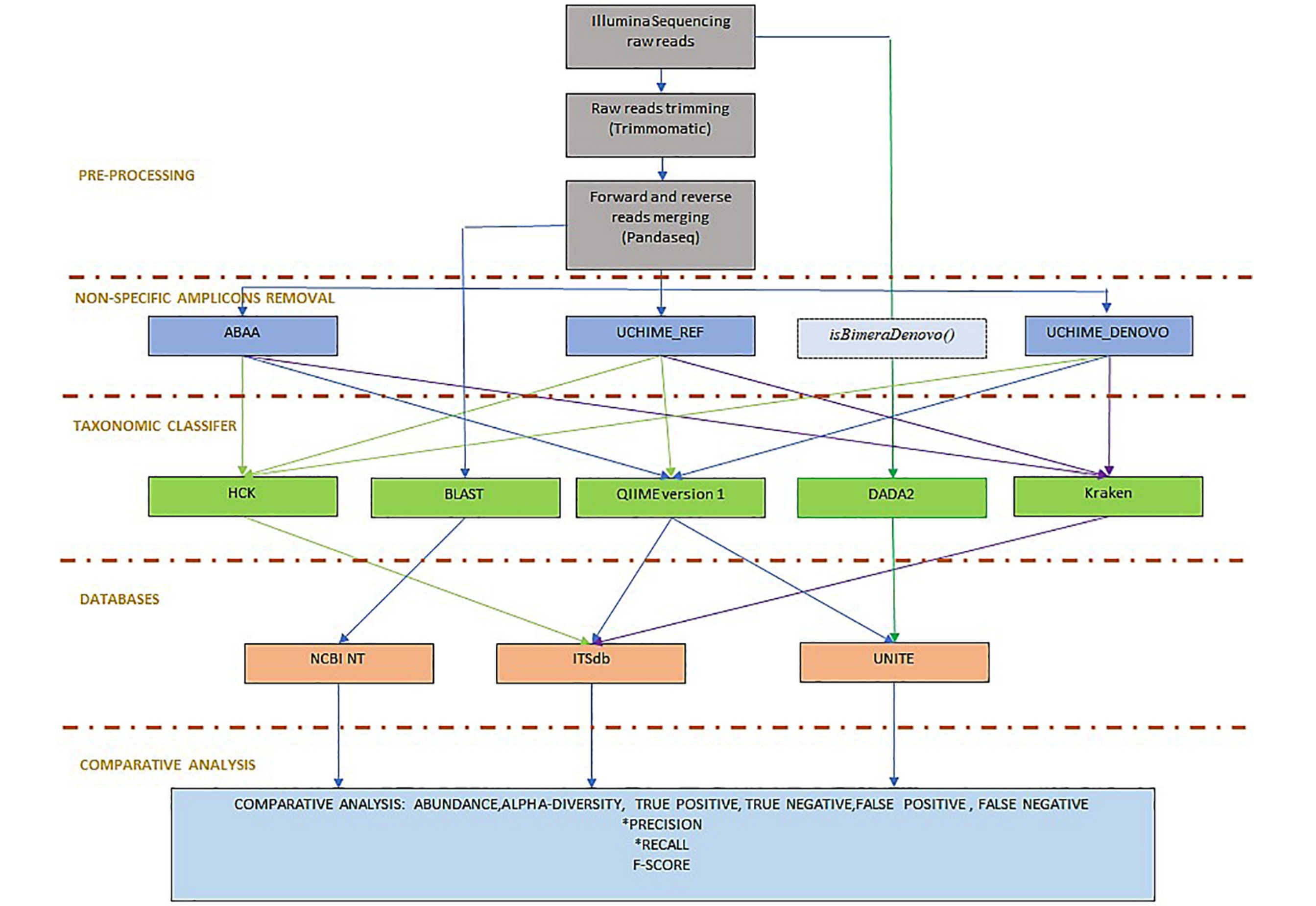

Illumina Sequencing raw reads
Raw reads trimming (Trimmomatic)
PRE-PROCESSING
Forward and reverse reads merging (Pandaseq)
NON-SPECIFIC AMPLICONS REMOVAL
ABAA
UCHIME_REF
isBimeraDenovo()
UCHIME_DENOVO
TAXONOMIC CLASSIFER
HCK
BLAST
QIIME version 1
DADA2
Kraken
DATABASES
NCBI NT
ITSdb
UNITE
COMPARATIVE ANALYSIS
COMPARATIVE ANALYSIS: ABUNDANCE,ALPHA-DIVERSITY, TRUE POSITIVE, TRUE NEGATIVE,FALSE POSITIVE , FALSE NEGATIVE
*PRECISION
*RECALL
F-SCORE

FIGURE 2 | Benchmarking workflow: the workflow is organised in pre-processing, including Illumina sequencing reads trimming and forward and reverse reads merging. To determine the best non-specific amplicons filtering method, ABAA was tested with UCHIME (UCHIME_Ref, UCHIME_DENOVO, *and isBimeraDenovo()* implemented in DADA2. The classifier tested also includes HCK with ITSdb, the newly designed pipeline, QIIME with ITSdb and UNITE, Kraken with ITSdb, DADA2 with UNITE and compared to BLASTn search using NCBI NT database, as reference. The pipelines performances are evaluated using precision, recall and F-score.

251 bp, and the minimum is 35 bp (**Figure 3A**). Fragments with a read length below 150 bp have fewer duplicated percentages than those between 240 and 250. After joining the paired reads, the average size is 233.3 bp (*SD* = 94.45) with a maximum of 472 bp and a minimum of 35 bp. The predominant amplicons size is 251 bps. We observed low-frequency fragments below 250 bp and above 400 bp (**Figure 3B**).

Taxonomic Assignment Using BLAST: Mock Communities Real Abundance Estimation as a Reference

We conducted a BLASTn search against the NCBI NT database to re-estimate the absolute abundance of the expected genus. We observed discrepancies between theoretical data and BLAST results. Even though samples SRR5439721 and SRR5439722 were from the same mock preparation, *Aspergillus* amplicons could not be detected in SRR5439721, and *Cryptococcus* was undermined in sample SRR5439721. *Malassezia* was also undermined in SRR6702280, SRR6702281, SRR6702283. The details of the abundance table are shown in **Table 1**.

Benchmark and Comparative Analysis: Performance of HCK and ABAA

ABAA: Amplicons Filtering

For our analysis, amplicons length below 250 bp and above 400 bp have shown low frequency compared to those comprised between 251 and 400 bp (**Figure 3D**). Each peak in **Figure 3C** is composed of amplicons of a similar size. The enlargement of the base of the curve may correspond to the variation of the amplicon's size. The frequency of these amplicons indicates that they could also be derived from non-specific amplification. Here we hypothesise this amplicon to be a chimaera and attempt to filter them out. The minimum sequence length detected by ABAA is 35 bp, with an average of 308 bp, higher than the overall average length (233 bp) and a maximum of 472 bp. It indicates that most chimaeras formed in this dataset may result from bimera and or multimera forming than

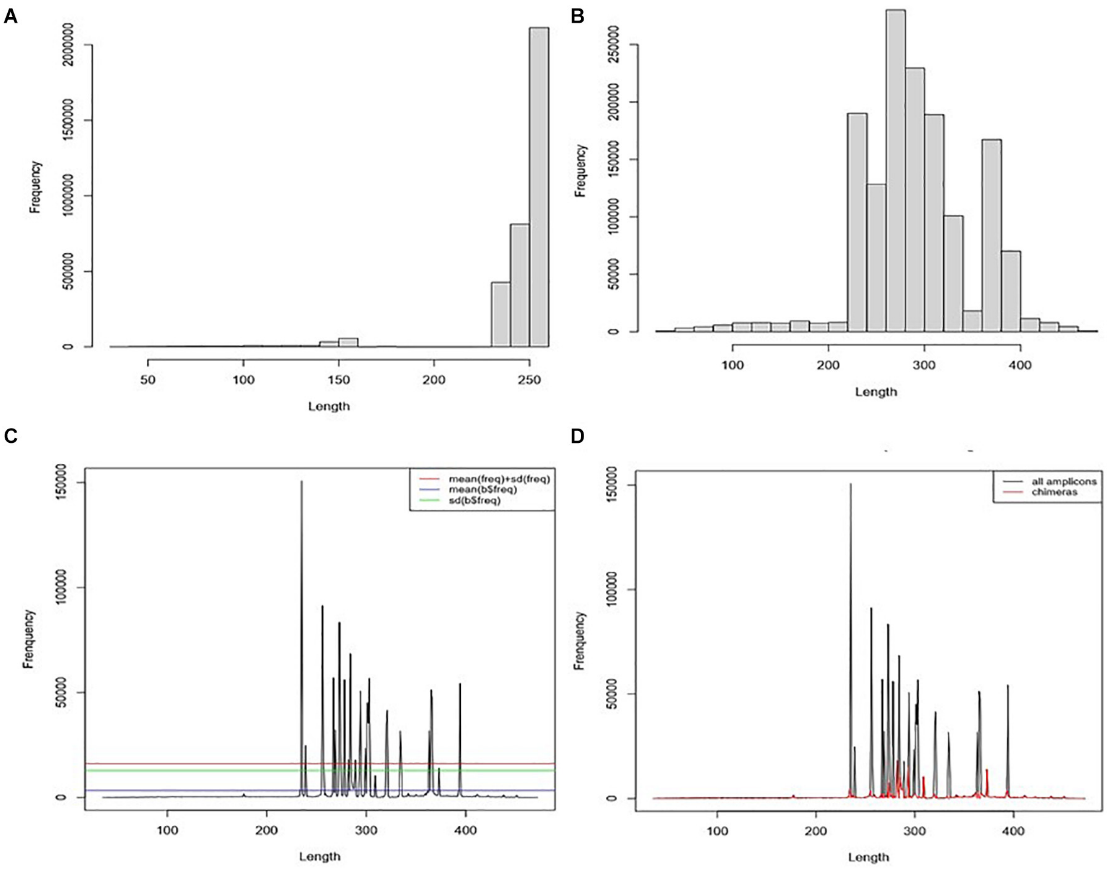


FIGURE 3 | Chimaera detection flow using *ABAA*: **(A)** distribution of reads from all datasets; The average read length is 200.9 bp (*SD* = 65.6, a median of 250), the maximum reads size is 251 bp, and the minimum is 35 bp. **(B)** Distribution of the frequency of contigs length (assembly of forward and reverse reads), the average size is 233.3 bp (*SD* = 94.45) with a maximum of 472 bp and a minimum of 35 bp. The predominant amplicons size is 251 bps. **(C)** Distribution of contigs length-frequency by length: determination of non-specific amplicons filtering cut-off: cut-off was tested for means (blue line), standard deviation (green line), and mean + standard deviation (red line). The standard deviation was kept for better performance. **(D)** Distribution of sequences by the length in all datasets. Blackline represents the distribution of all sequences). Moreover, the red line represents the distribution of filtered sequences with ABAA (standard deviation).

incomplete amplification. Filtered amplicons by ABAA include amplicons below 250 bp, above 400 bp, and low amplification between 250 and 400 bp (**Figure 3D**). In total ABAA has detected 252,567 sequences accounting for 10.86% of overall sequences as non-specific amplicons while 528,544 (22.72%) with uchime_ref and 1,165,031 (50.08%) by DADA2 and 32 (0.0013%) detected by uchime_denovo. *isBimeraDenovo() in* DADA2 has filtered out up to 75.43% of sequences in sample SRR5439722. However, 24.76% were detected with UCHIME_REF, and 17.02% by ABAA on the other hand. Also, 51.74% were detected in sample SRR5439721, while 24.67% detected with UCHIME_REF and 17.23% by ABAA and UCHIME_REF seem to be more consistent than *isBimeraDenovo()* in DADA2 as samples SRR5439721 and SRR5439722 were from the same mock preparation (**Table 2**).

HCK-ABAA: Taxonomic Assignment Performances

The second step of the HCK pipeline handles the chimaeras-free sequences. Samples sequences pre-processed and filtered by ABAA in the first step are then combined into a single fasta for the clustering process. With our datasets, we cluster a total of 2,326,239 amplicons with HCK using multiple-step iterated runs of cd-hit-est to perform hierarchical clustering. The first iteration performed with 99% sequence similarity generates 88,428 clusters, the second iteration with 98% creates 32,200 clusters (3/8 of the initial clusters), and the final iteration at 97% produces 18,831 clusters. The final iteration reduces the total clusters by 1/5 of the initial clusters, a crucial benefit of the hierarchical clustering that will be detailed in the discussion. All clusters generated by different iterations are merged into 18,770 non-redundant clusters, including 14,431 singletons, for which 2,545 have fragments size $\leq$ 149 bp and 11,886 with sequence size $\geq$150 bp (150 bp, widely considered as the minimum standard of ITS length). The singletons are removed from further analysis based on the assumption that a unique sequence might derive from sequencing errors or non-specific amplification. As a result, only 4,339 clusters are composed of biologically valid amplicons corresponding to 4,339 representative sequences. The performance of HCK with and without ABAA is assessed using the precision (positive predictive value), the recall (sensitivity),

TABLE 1 | Absolute count of reads affiliated of each genus among the different datasets of the mock community (determined by BLASTn against NT database).

Taxa	SRR 5439721	SRR 5439722	SRR 6702280	SRR 6702281	SRR 6702283	SRR 8473974	SRR 8473977	SRR 8473978	SRR 8473979	SRR 8473980	SRR 8473984
Alternaria	0	0	3,941	3,334	5,462	39,009	0	0	0	0	0
Apophysomyces	0	0	0	0	0	0	0	0	0	0	191
Aspergillus	0	1,381	924	1,029	1,120	315,986	5,991	5,096	5,657	3,613	2,362
Blastomyces	0	0	0	0	0	0	379	230	256	183	0
Candida	150,704	108,934	33,556	28,123	39,320	125,949	0	0	0	0	0
Cladosporium	0	0	51	66	76	22,476	0	0	0	0	0
Clavispora	0	0	0	0	0	1,645	0	0	0	0	0
Coccidioides	0	0	0	0	0	0	502	649	698	515	0
Cryptococcus	1,087	33,983	12	9	16	16	555	568	825	665	131
Cunninghamella	0	0	0	0	0	0	0	0	0	0	13
Exophiala	0	0	3,340	2,791	4,140	0	299	271	319	221	2
Fusarium	0	0	52	54	76	39,505	0	0	0	0	138
Histoplasma	0	0	0	0	0	0	152	112	138	87	0
Lichtheimia	0	0	0	0	0	0	0	0	0	0	45
Malassezia	0	0	50	67	67	0	0	0	0	0	0
Meyerozyma	0	0	0	0	0	15,220	0	0	0	0	0
Mucor	0	0	0	0	0	11,935	109	96	109	80	438
Paecilomyces	0	0	0	0	0	0	0	0	0	0	80
Penicillium	0	0	32,037	29,551	39,368	27,888	1,645	1,461	1,593	854	0
Phanerochaete	32,650	20,763	0	0	0	0	0	0	0	0	0
Pichia	22,965	69,159	0	0	0	95,622	0	0	0	0	0
Purpureocillium	0	0	0	0	0	0	120	120	128	96	0
Rasamsonia	0	0	2	0	0	292	0	0	0	0	128
Rhizopus	0	0	0	0	0	57	2,283	2,354	3,447	2,584	109
Saccharomyces	57,107	46,582	1,737	1,863	1,942	12,654	0	0	0	0	0
Sarocladium	0	0	0	0	0	0	257	269	336	221	0
Scedosporium	0	0	0	0	0	3	62	75	86	54	36
Schizosaccharomyces	36,413	23,880	0	2	0	7	0	0	0	0	0
Sporidiobolus	0	0	0	0	0	17,476	0	0	0	0	0
Sporothrix	0	0	0	0	0	2	0	0	0	0	1
Talaromyces	0	0	0	0	0	1,336	435	382	403	160	6
Trichoderma	29,027	18,431	0	0	0	0	0	0	0	0	0
Trichosporon	0	0	122	77	185	88,790	2,305	2,080	2,221	1,642	0
Wickerhamomyces	2,208	1,235	0	0	0	10	0	0	0	0	0
Sclerotina	25,545	12,058	0	0	0	0	0	0	0	0	0
Rhyzomucor	0	0	0	0	0	0	0	0	0	0	0
Trichophyton	0	0	7,142	6,084	9,087	0	0	0	0	0	0

and the F-score based on the true-positive, false-positive, and false-negative rates as described in material and method. This performance is compared to other pipelines, e.g., QIIME version 1 with both databases ITSdb and UNITE, Kraken version1 with ITSdb, and DADA2 with UNITE database. All classification methods are tested with and without the chimaera removal step. The analysis shows that HCK without non-specific amplicons removal is slightly better than Kraken (precision: 0.685 and 0.682, recall: 0.986 and 0.983 and F-score: 0.80 and 0.79, respectively) and HCK decreases by 13.22% the false-positive detection and by 45.36% of false negatives compared to Kraken. The chimaera removal step with UCHIME_REF reduces the false positives by 32.05% and the false-negative by 10.44% compared to raw sequence processing. However, adding a step of chimaera filtering affects the sensitivity (recall), regardless of the method (**Table 3**).

HCK yields better classification performance when ABAA is added upstream (precision 0.83 and the second best is HCK/UCHIME_REF with 0.8), and consequently, the F-score is also improved (0.89, **Figure 4**). Besides, the association of HCK and ABAA reduces the proportion of false-positive by 35.52% compared to HCK with UCHIME_REF and 97.01% without non-specific sequences removal. QIIME used with UCHIME_REF, and ITSdb performs better (F-score = 0.716) than similar approach with UNITE database (f-score = 0.648). DADA2 was also tested with its filtering method includes in the pipeline. The true-positive sequence classified was shallow compared to others, and this might be due to the high number of chimaeras

TABLE 2 | Level of detection of chimaera removal methods.

Samples	Total sequences	Chimaera ABaa	Chimaera free Abaa*	Chimaera uchime_ref	Chimaera free uchime_ref*	Chimaera uchime_denovo	Chimaera free uchime_denovo*	Chimaera dada2	Chimaera free dada2*
SRR5439721	426,354	73,482 (17.23%)	352,872	105,167 (24.67%)	321,187	03 (00007%)	426,351	220,607 (51.74%)	205,747
SRR5439722	397,268	68,091 (17.02%)	329,177	98,367 (24.76%)	298,901	00 (0%)	397,268	299,665 (75.43%)	97,603
SRR6702280	174,185	19,187 (11.02%)	154,998	61,937 (35.56%)	112,248	11 (0.0063%)	174,174	71,129 (40.84%)	103,056
SRR6702281	148,397	17,980 (12.12%)	130,417	55,059 (37.10)	93,338	09 (0%)	148,388	63,247 (42.62%)	85,150
SRR6702283	246,368	31,616 (12.83%)	214,752	90,865 (36.88%)	155,503	08 (0.0032%)	246,360	93,543 (37.97%)	152,825
SRR8473974	865,248	40,730 (4.71%)	824,518	112,068 (12.95%)	753,180	01 (00001%)	865,247	388,352 (44.88%)	476,896
SRR8473977	17,181	379 (2.21%)	16,802	1,566 (9.11%)	15,615	00 (0%)	17,181	7,044 (41.00%)	10,137
SRR8473978	15,694	121 (0.77%)	15,573	1,363 (8.68%)	14,331	00 (0%)	15,694	6,510 (41.48%)	9,184
SRR8473979	18,394	295 (1.60%)	18,099	1,336 (7.26%)	17,058	00 (0%)	18,394	7,421 (40.34%)	10,973
SRR8473980	12,730	676 (5.31%)	12,054	387 (3.04%)	12,343	00 (0%)	12,730	4,898 (38.48%)	7,832
SRR8473984	4,420	10 (0.23%)	4,410	429 (9.71%)	3,991	00 (0%)	4,420	2,615 (59.16%)	1,805
Total	2,326,239	252,567 (10.86%)	2,073,672	528,544 (22.72%)	1,797,695	32 (0.0013%)	2,326,207	1,165,031 (50.08%)	1,161,208

**Sequence filtered with the corresponding method and cleaned.*

TABLE 3 | Precision, recall, accuracy, and F-score performance of HCK and ABAA version other tested combination of chimaera detection and taxonomy assignment methods.

Taxa. assign.	Chimaera remov.	Database	Total sequences	Unclassified sequences	True positive	False positive	False negative	Precision	Recall	F-score
hck	ABaa	ITSdb	2,073,672	378046	1,528,646	166,980	27,314	**0.834998**	0.97222	**0.89594**
hck	Uchime_ref	ITSdb	1,797,695	231779	1,306,938	258,978	249,022	0.803725	0.84344	0.81431
hck	Uchime_denovo	ITSdb	2,326,207	102360	1,547,646	676,201	8,314	0.670532	0.98645	0.79213
hck	–	ITSdb	2,326,239	220025	1,547,649	558,565	8,311	0.685258	**0.98645**	0.8013
kraken	ABaa	ITSdb	2,073,672	73480	1,497,681	502,511	58,279	0.713001	0.9651	0.81431
kraken	Uchime_ref	ITSdb	1797695	138673	1,277,882	381,140	278,078	0.725327	0.83483	0.77351
kraken	Uchime_denovo	ITSdb	2,326,207	141772	1,540,747	643,688	15,213	0.682455	0.98295	0.7977
kraken	–	ITSdb	2,326,239	141804	1,540,747	643,688	15,213	0.682455	0.98295	0.7977
qiime	ABaa	ITSdb	2,073,672	407960	1,369,677	296,035	186,283	0.782386	0.83022	0.78878
qiime	Uchime_ref	ITSdb	1,797,695	389342	1,175,806	232,547	380,154	0.765336	0.72615	0.71648
qiime	Uchime_denovo	ITSdb	2,326,207	515687	1,391,958	418,562	164,002	0.734174	0.83587	0.76968
qiime	–	ITSdb	2,326,239	525891	1,391,597	408,751	164,363	0.735705	0.83605	0.77068
qiime	ABaa	Unite	2,073,672	455026	1,208,721	409,925	347,239	0.529577	0.76841	0.62466
qiime	Uchime_ref	Unite	1,797,695	43024	1,050,365	704,306	505,595	0.510642	0.64894	0.56654
qiime	Uchime_denovo	Unite	2,326,207	347217	1,310,235	668,755	245,725	0.527798	0.79978	0.63181
qiime	–	Unite	2,326,239	52075	1,307,126	967,038	248,834	0.539557	0.79841	0.6396
Dada2		Unite	1,161,208	0	750,588	410,620	805,372	0.717852	0.5755	0.62471

These values were highlighted in bold to show they are the top values.

sequences filtered (50.08%). Its F-score performance is 0.62, with 411 775 false-positive and 805 372 false negatives. The Kraken based classification implemented with chimaera methods yields comparable sensitivity results (recall) to that of HCK, but the higher number of false-positive impacts the precision and the overall performance (F-score: 0.79) (**Table 3**).

Diversity Metrics Analysis

One of the most significant endpoints of ITS sequencing is the comparison of alpha diversity; thus, we compare the alpha diversity of all tested classification methods to that of BLASTn using Chao1 and Shannon indexes, assuming that diversity with BLAST search is closer to reality. With the Chao1 index, HCK diversity is closed to BLASTn estimation compared to Kraken, QIIME, and DADA2 estimation. With chao1, HCK in association with ABAA held the lowest difference with BLASTn (54.02), followed by HCK with UCHIME_DENOVO. With the Shannon index, HCK, used with UCHIME_REF, held the best rank(0.76), followed by HCK with ABAA (**Supplementary Table 1**). The data show that the BLASTn search estimates the chao1 index between 8 and 14 for all the samples. HCK with ABAA chaos1estimates is between 20 and 80, and kraken with and without chimaera removal's estimation between 176 and 856. DADA2's chao1 estimation is also close to the BLASTn search's; however, there is an overestimation for some samples (15−650, **Figure 5**). The average Shannon index diversity of BLASTn search is 2 for all samples. It varies between 2 and 3 with HCK and ABAA, 2−7 for DADA2, 3 for Kraken with or without chimaera removal, and up to 6 for QIIME, depending on the database and the chimaera removal method (**Figure 5**).

Computing Specification and Speed

ABAA and HCK computing resources were tested and timed on ubuntu-based system 20.04, WSL2; Processor: Intel® Core^{TM} i7-8650U CPU @ 1.90GHz 2.11 GHz; RAM 16.0 GB; System type 64-bit Operating System, x64-based processor. The running speed for each method tested is listed in **Supplementary Table 2**.

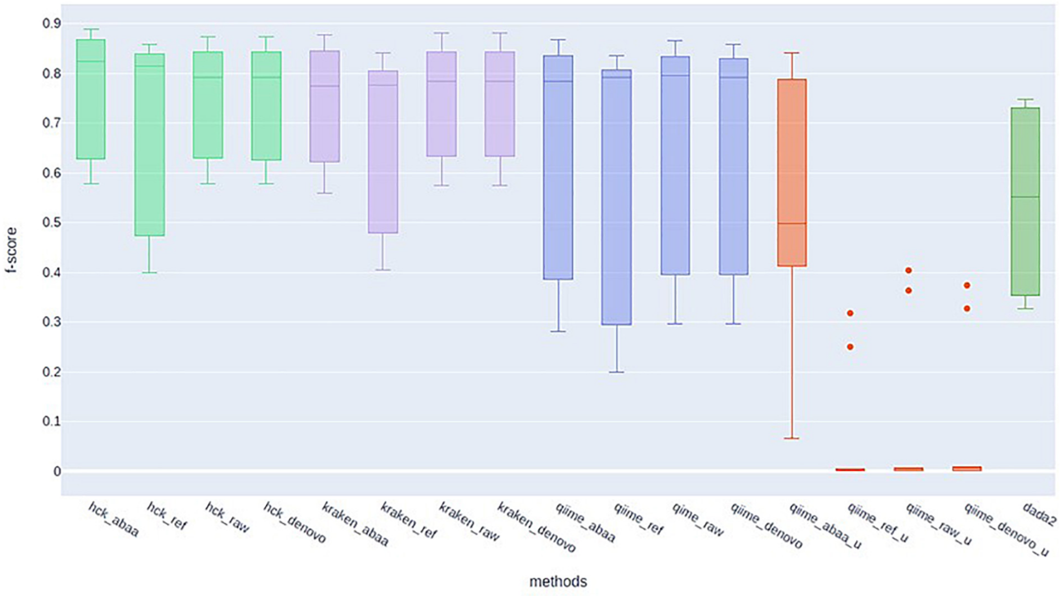


FIGURE 4 | F-score performance of HCK (light green), Kraken (mauve), Qiime with ITSdb (blue), Qiime with Unite database (red), and DADA2 (green).

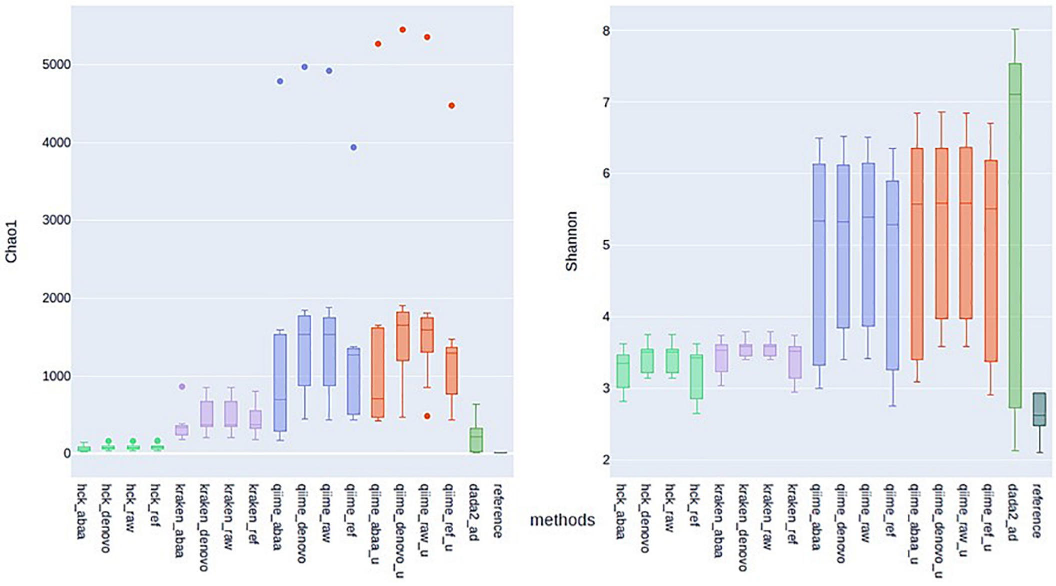


FIGURE 5 | Estimation of Alpha diversity metric with Chao1 (left) and Shannon (right) indexes calculated with the estimated abundance of different methods and association with various chimaera removal methods

ABaa filtered 2,326,239 sequences (623 MiB) in less than 2 min (1 min 18 s), while this requires almost 50 min for UCHIME_REF. *isBimeraDenovo()* function of DADA2 could not be tested separately as this step depends on many other steps. HCK and Qiime v1 process 23,700 sequences per min on average, while DADA2 processes 65,300 sequences per min. This speed does not include sequence truncation as it is not required for ITS processing.

DISCUSSION

In this work, we present full fungi ITS based classification workflow using two newly developed tools, ABAA and HCK, to filter non-specific amplicons and taxonomically classify them. We compare the performance of ABAA to UCHIME and *isBimeraDenovo() in* DADA2 and the performance of HCK to that of Kraken, QIIME, and DADA2 using publicly available mock community Illumina sequencing datasets. The analysis revealed that HCK-ABAA yields the best performance. The F-score is systematically improved when ABAA is used to filter amplicons regardless of the classifier. This work also shows the impact of filtering methods on the ecological diversity metrics and how they can dramatically change the estimation of a sample's diversity.

The efficiency of sequence amplification and the quality of sequencing reads are critical and determinants for the outcome of the metabarcoding analysis and especially for fungi ITS locus (Schloss et al., 2011). Non-specific amplicons present a serious threat to the classification and taxa abundance estimation. The size and the number of ITS loci are highly variable, unlike the 16S rRNA gene in bacteria and sufficiently polymorphic to delineate fungi at the genus and or species level (Tang et al., 2015; Khodadadi et al., 2017). This variation can be biological or can derive from high rates of insertions and deletions in the evolution of this less conserved genetic region. It can also derive from non-specific amplification. ABAA has this advantage of considering the real distribution of amplicons size from real datasets. It does not need database maintenance and only requires minimum computing resources. It filters sequences based on the distribution of their size-frequency and mainly targets amplicons with low length-frequency. The performance of the majority of chimaera filtering methods are usually assessed on simulated chimaera sequences (Nilsson et al., 2010; Harris et al., 2012), but when applied to the real dataset, it is challenging to determine whether sequences that have been filtered are real chimaeras. The fragment size dissimilarity also creates bias during conventional clustering. Consequently, this affects OTUs picking and abundance estimation, including overestimating or underestimating community abundance (De Filippis et al., 2017). Except for Kraken, the majority of metabarcoding methods include a clustering process. Clustering consists of reducing the amplicons similarity redundancy of data diversity. The most commonly used in amplicon metabarcoding analysis are uclust in the usearch algorithm (Edgar, 2010), vsearch (Rognes et al., 2016), and CD-HIT (Fu et al., 2012). usearch and vsearch can cluster nucleotide sequences based on their similarity, length, and abundance, assuming that the same species' amplicons will probably be identical in size with a minimal coverage dissimilarity. As a result, with fungi ITS, clustering may create multiple OTUs from the same species amplicons and increase the alpha and the beta diversities.

CD-HIT implements a more realistic clustering approach, hierarchical clustering, which consists of a multiple-step, iterated runs with a neighbour-joining approach and generates a hierarchical structure. In HCK, with the datasets that we analyse, the second iteration with 98% identity reduces the first number of clusters by 3/8 and the final iteration with 97% identity by 1/5. In addition to filtering out the singleton, HCK drastically reduces the number of false-positive and normalises the diversity abundance. It is essential to highlight that databases also play an important role in the performance of the classifier. Qiime version 1 performs better with the ITSdb database than its native database UNITE, regardless of the filtering method. The inappropriate estimation of the abundance (overestimation, underestimation of population or sequence wrongly classified) can also influence metrics of diversity. The high diversity found with the UNITE database might be due to the higher number of incorrect classification sequences in the UNITE database.

CONCLUSION

The classification of fungi using ITS marker is very challenging. It is owed to the high diversity of the kingdom. Moreover, targeting an intergenic section as ITS1 leads to diversified amplicon sizes and sequences that are not taken into account with the classical approaches developed for 16S analysis. Combining HCK and ABAA increases the number of true-positive and decreases the proportion of false-positive, as shown with the datasets we have evaluated. Consequently, HCK maintained the alpha diversity metric with the Chao1 index close to that of the BLASTn, compared to QIIME, Kraken, and DADA2. As demonstrated in this analysis, the use of HCK in association with ABAA allows a more realistic estimation of fungal diversity. So far, it is the best option to perform fungi ITS1 metabarcoding analysis on clinical and non-clinical samples.

DATA AVAILABILITY STATEMENT

Sequences analysed in this project are available under accession numbers SRR8473974, SRR8473977, SRR8473978, SRR8473979, SRR8473980, SRR8473984, SRP132544, SRR6702280, SRR6702281, SRR6702283, SRR5439721, and SRR5439722 in NCBI. The original contributions presented in the study are included in the Material and method section, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

AD and OP supervised the study. KM analysed interpreted data and wrote the manuscript. CB, AM, AO, and AK made critical revisions. All authors read and approved the final manuscript.

FUNDING

L'Oréal Research and Innovation Chair supported this work.

ACKNOWLEDGMENTS

We thank Antoine Bodein and Simon Gotty for their valuable contribution toward the scripts writing and improvement, as well as Léo D'agata for his comments. We also thank Thomas Jeanne and Wen Chen for their review.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.640693/full#supplementary-material

REFERENCES

Abarenkov, K., Henrik Nilsson, R., Larsson, K.-H., Alexander, I. J., Eberhardt, U., Erland, S., et al. (2010a). The UNITE database for molecular identification of fungi - recent updates and future perspectives. *New Phytol.* 186, 281–285. doi: 10.1111/j.1469-8137.2009.03160.x

Abarenkov, K., Tedersoo, L., Nilsson, R. H., Vellak, K., Saar, I., Veldre, V., et al. (2010b). Plutof-a web-based workbench for ecological and taxonomic research, with an online implementation for fungal its sequences. *Evol. Bioinform.* 6, 189–196. doi: 10.4137/EBO.S6271

Bazzicalupo, A. L., Bálint, M., and Schmitt, I. (2013). Comparison of ITS1 and ITS2 rDNA in 454 sequencing of hyperdiverse fungal communities. *Fungal Ecol.* 6, 102–109. doi: 10.1016/j.funeco.2012.09.003

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., and Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiol.* 10:189. doi: 10.1186/1471-2180-10-189

Bjørnsgaard Aas, A., Davey, M. L., and Kauserud, H. (2017). ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Mol. Ecol. Resour.* 17, 730–741. doi: 10.1111/1755-0998.12622

Blaalid, R., Kumar, S., Nilsson, R. H., Abarenkov, K., Kirk, P. M., and Kauserud, H. (2013). ITS1 versus ITS2 as DNA metabarcodes for fungi. *Mol. Ecol. Resour.* 13, 218–224. doi: 10.1111/1755-0998.12065

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

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., et al. (2018). QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. *Nat Biotechnol.* 37, 852–857. doi: 10.7287/peerj.preprints.27295v1

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

De Filippis, F., Laiola, M., Blaiotta, G., and Ercolini, D. (2017). Different amplicon targets for sequencing-based studies of fungal diversity. *Appl. Environ. Microbiol.* 83:e00905-17. doi: 10.1128/AEM.00905-17

Edgar, R. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. *bioRxiv* [Preprint] doi: 10.1101/074252

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Fosso, B., Santamaria, M., Marzano, M., Alonso-Alemany, D., Valiente, G., Donvito, G., et al. (2015). BioMaS: a modular pipeline for bioinformatic analysis of metagenomic AmpliconS. *BMC Bioinform.* 16:203. doi: 10.1186/s12859-015-0595-z

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT : accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Fujita, S. I., Senda, Y., Nakaguchi, S., and Hashimoto, T. (2001). Multiplex PCR using internal transcribed spacer 1 and 2 regions for rapid detection and identification of yeast strains. *J. Clin. Microbiol.* 39, 3617–3622. doi: 10.1128/JCM.39.10.3617-3622.2001

Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E., et al. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ.* 7:e6160. doi: 10.7717/peerj.6160

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., et al. (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods Ecol. Evol.* 6, 973–980. doi: 10.1111/2041-210X.12399

Harris, S. R., Clarke, I. N., Seth-Smith, H. M. B., Solomon, A. W., Cutcliffe, L. T., Marsh, P., et al. (2012). Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nat. Genet.* 44, 413–419, S1. doi: 10.1038/ng.2214

Hoggard, M., Vesty, A., Wong, G., Montgomery, J. M., Fourie, C., Douglas, R. G., et al. (2018). Characterising the human mycobiota: a comparison of small subunit rRNA, ITS1, ITS2, and large subunit rRNA genomic targets. *Front. Microbiol.* 9:2208. doi: 10.3389/fmicb.2018.02208

Khodadadi, H., Karimi, L., Jalalizand, N., Adin, H., and Mirhendi, H. (2017). Utilisation of size polymorphism in ITS1 and ITS2 regions for identification of pathogenic yeast species. *J. Med. Microbiol.* 66, 126–133. doi: 10.1099/jmm.0.000426

Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., et al. (2017). Deciphering diversity indices for a better understanding of microbial communities. *J. Microbiol. Biotechnol* 27, 2089–2093. doi: 10.4014/jmb.1709.09027

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–5277. doi: 10.1111/mec.12481

Kumar, S., Carlsen, T., Mevik, B. H., Enger, P., Blaalid, R., Shalchian-Tabrizi, K., et al. (2011). CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinform.* 12:182. doi: 10.1186/1471-2105-12-182

Lahr, D. J. G., and Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47, 857–866. doi: 10.2144/000113219

Martin, K. J., and Rygiewicz, P. T. (2005). Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. *BMC Microbiol.* 5:28. doi: 10.1186/1471-2180-5-28

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13:31. doi: 10.1186/1471-2105-13-31

McTaggart, L. R., Copeland, J. K., Surendra, A., Wang, P. W., Husain, S., Coburn, B., et al. (2019). Mycobiome sequencing and analysis applied to fungal community profiling of the lower respiratory tract during fungal pathogenesis. *Front. Microbiol.* 10:512. doi: 10.3389/fmicb.2019.00512

Mysara, M., Vandamme, P., Props, R., Kerckhof, F. M., Leys, N., Boon, N., et al. (2017). Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiol. Ecol.* 93:fix029. doi: 10.1093/femsec/fix029

Nilsson, R. H., Abarenkov, K., Veldre, V., Nylinder, S., de Wit, P., Brosché, S., et al. (2010). An open source chimera checker for the fungal ITS region. *Mol. Ecol. Resour.* 10, 1076–1081. doi: 10.1111/j.1755-0998.2010.02850.x

Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E., Sjökvist, S., and Kristiansson, E. (2008). The ITS region as a target for characterisation of fungal communities using emerging sequencing technologies. *FEMS Microbiol. Lett.* 296, 97–101. doi: 10.1111/j.1574-6968.2009.01618.x

Ninet, B., Jan, I., Bontems, O., Léchenne, B., Jousson, O., Panizzon, R., et al. (2003). Identification of dermatophyte species by 28S ribosomal DNA sequencing with a commercial kit. *J. Clin. Microbiol.* 41, 826–830. doi: 10.1128/JCM.41.2.826-830.2003

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the {Effects} of {PCR} {Amplification} and {Sequencing} {Artifacts} on 16S {rRNA}-{Based} {Studies}. *PLoS One* 6:e27310. doi: 10.1371/journal.pone.0027310

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Tang, J., Iliev, I. D., Brown, J., Underhill, D. M., and Funari, V. A. (2015). Mycobiome: approaches to analysis of intestinal fungi. *J. Immunol. Methods* 421, 112–121. doi: 10.1016/j.jim.2015.04.004

Wang, X. C., Liu, C., Huang, L., Bengtsson-Palme, J., Chen, H., Zhang, J. H., et al. (2015). ITS1: a DNA barcode better than ITS2 in eukaryotes? *Mol. Ecol. Resour.* 15, 573–586. doi: 10.1111/1755-0998.12325

White, J. R., Maddox, C., White, O., Angiuoli, S. V., and Fricke, W. F. (2013). CloVR-ITS: automated internal transcribed spacer amplicon sequence analysis pipeline for the characterisation of fungal microbiota. *Microbiome* 1:6. doi: 10.1186/2049-2618-1-6

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R45 doi: 10.1186/gb-2014-15-3-r46

Wu, S., Xiong, J., and Yu, Y. (2015). Taxonomic resolutions based on 18S rRNA genes: a case study of subclass copepoda. *PLoS One* 10:e0131498. doi: 10.1371/journal.pone.0131498

Zajec, N., Stres, B., and Avguštin, G. (2012). Distinct approaches for the detection and removal of chimeric 16S rRNA sequences can significantly affect the outcome of between-site comparisons. *Aquat. Microb. Ecol.* 66, 13–21. doi: 10.3354/ame01510

Conflict of Interest: AO, AK and OP were employed by the company L'Oreal.

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.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fmicb.2021.680101

Dysbiosis of Gut Microbiota in Patients With Acute Myocardial Infarction

Ying Han[1], Zhaowei Gong[1], Guizhi Sun[1], Jing Xu[1], Changlu Qi[2], Weiju Sun[3], Huijie Jiang[4]*, Peigang Cao[5]* and Hong Ju[6]*

[1] Department of Cardiovascular, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [3] Department of Cardiovascular, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [4] Department of Radiology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [5] Department of Cardiology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China, [6] Department of Information Engineering, Heilongjiang Biological Science and Technology Career Academy, Harbin, China

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Xuan Li,
University of Mississippi Medical Center, United States
Jiajie Peng,
Northwestern Polytechnical University, China

***Correspondence:**
Huijie Jiang
jianghuijie@hrbmu.edu.cn
Peigang Cao
caopeigang@126.com
Hong Ju
hongju.hit@hotmail.com

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

Received: 13 March 2021
Accepted: 17 May 2021
Published: 05 July 2021

Citation:
Han Y, Gong Z, Sun G, Xu J, Qi C, Sun W, Jiang H, Cao P and Ju H (2021) Dysbiosis of Gut Microbiota in Patients With Acute Myocardial Infarction. Front. Microbiol. 12:680101. doi: 10.3389/fmicb.2021.680101

Acute myocardial infarction (AMI) continues as the main cause of morbidity and mortality worldwide. Interestingly, emerging evidence highlights the role of gut microbiota in regulating the pathogenesis of coronary heart disease, but few studies have systematically assessed the alterations and influence of gut microbiota in AMI patients. As one approach to address this deficiency, in this study the composition of fecal microflora was determined from Chinese AMI patients and links between gut microflora and clinical features and functional pathways of AMI were assessed. Fecal samples from 30 AMI patients and 30 healthy controls were collected to identify the gut microbiota composition and the alterations using bacterial 16S rRNA gene sequencing. We found that gut microflora in AMI patients contained a lower abundance of the phylum *Firmicutes* and a slightly higher abundance of the phylum *Bacteroidetes* compared to the healthy controls. Chao1 ($P = 0.0472$) and PD-whole-tree ($P = 0.0426$) indices were significantly lower in the AMI versus control group. The AMI group was characterized by higher levels of the genera *Megasphaera*, *Butyricimonas*, *Acidaminococcus,* and *Desulfovibrio*, and lower levels of *Tyzzerella 3*, *Dialister*, *[Eubacterium] ventriosum group*, *Pseudobutyrivibrio*, and *Lachnospiraceae ND3007 group* as compared to that in the healthy controls ($P < 0.05$). The common metabolites of these genera are mostly short-chain fatty acids, which reveals that the gut flora is most likely to affect the occurrence and development of AMI through the short-chain fatty acid pathway. In addition, our results provide the first evidence revealing remarkable differences in fecal microflora among subgroups of AMI patients, including the STEMI vs. NSTEMI, IRA-LAD vs. IRA-Non-LAD and Multiple (≥2 coronary stenosis) vs. Single coronary stenosis groups. Several gut microflora were also correlated with clinically significant characteristics of AMI patients, including LVEDD, LVEF, serum TnI and NT-proBNP, Syntax score, counts of leukocytes, neutrophils and monocytes, and fasting serum glucose levels. Taken together, the data generated enables the prediction of several functional pathways as based on the fecal microfloral composition of AMI patients. Such information may enhance our comprehension of AMI pathogenesis.

Keywords: acute myocardial infarction, gut microflora, 16S rRNA gene, patients, dysbiosis

INTRODUCTION

Despite timely reperfusion through primary percutaneous coronary intervention, acute myocardial infarction (AMI) is still the leading cause of morbidity and mortality worldwide (Nwokocha et al., 2017; Khan et al., 2020; Qi et al., 2021). It is really important to further reduce myocardial infarct size and preserve cardiac function, in order to reduce risk of death and prevent onset of heart failure. Therefore, it is necessary to thoroughly understand its pathogenic mechanism and find new therapeutic targets. Accumulating evidence reveals the role of gut microbiota in regulating the pathogenesis of coronary heart disease (CHD) (Cui et al., 2017; Liu et al., 2019, 2020; Moraru et al., 2019; Heianza et al., 2020; Marzullo et al., 2020), including the obvious association between the gut microbiota and the severity of AMI in rats (Lam et al., 2012, 2016; McCafferty et al., 2012).

There are a vast array of microbes in human gut, collectively referred to as the microbiota, which is a complex community. The metabolic activities and interactions with the immune system of gut microbiota are not limited the gut itself (Sonnenburg and Backhed, 2016), but also involve a variety of immune-mediated diseases and metabolic diseases such as diabetes, obesity, digestive system diseases, asthma, arthritis, cancers, and cardiovascular disease (Vieira et al., 2014; Sonnenburg and Backhed, 2016; Grigoryan et al., 2019; Ingham et al., 2019; Kolodziejczyk et al., 2019; Alemao et al., 2020; Morel et al., 2020; Yang et al., 2021). The abundance of *Enterobacteriaceae* and *Streptococcus* spp. were reported to be higher in patients with atherosclerotic cardiovascular disease compared to the healthy controls in a metagenome-wide association study (Jie et al., 2017). Koren et al. (2011) revealed that *Chryseomonas*, *Veillonella,* and *Streptococcus* exsited in AS plaque samples, and several bacterial flora in the intestine are the same as atherosclerotic plaques. Furthermore, they were related to the cholesterol levels (Koren et al., 2011). Trimethylamine-*N*-oxide (TMAO), a metabolite of gut microbiota, can partially promote the formation of atherosclerosis by promoting the formation of macrophage foam cells (Koeth et al., 2013; Tang et al., 2013). According to the study of Emoto et al. (2016) the alterations of gut microbiota were linked to the incidence of coronary artery disease.

Published studies on the role of the microbiome in coronary heart disease have investigated most patients with stable coronary heart disease using a cross-sectional design. However, there are few clinically meaningful prospective studies of the microbiome in patients with AMI. Therefore, in this study, fecal samples from AMI patients and healthy controls were collected, variable regions of gut bacterial 16S rRNA were amplified, and DNA library was constructed. The data of this study may provide detailed information on variations of gut microbial composition and its impacts on AMI.

MATERIALS AND METHODS

Study Participants

Between April and August of 2020, we recruited 30 AMI in-hospital patients and 30 asymptomatic controls receiving routine physical examinations for this study. AMI patients were recruited from the First and Fourth Affiliated Hospitals of Harbin Medical University while the healthy controls were recruited from the Fourth Affiliated Hospital of Harbin Medical University. AMI diagnosis was based on the World Health Organization definition and the third universal definition of myocardial infarction and consisted of patients with symptoms of ischemia, cardiac laboratory biomarker data, electrocardiogram results and invasive coronary angiograms or coronary CT angiography (Mendis et al., 2011; Thygesen et al., 2012). The criteria for the healthy controls included no ischemic symptoms, normal electrocardiogram and coronary stenosis of <25% as assessed by invasive coronary angiograms or coronary CT angiography. The exclusion criteria consisted of subjects that: (i) received antacids, probiotics, antibiotics, or antimicrobial agents within 30 days before sample collection, (ii) had an organic disease of the digestive system, and (iii) had gastrointestinal surgery. All participants experienced a normal lifestyle prior to admission, including typical Chinese diets based on carbohydrates versus high-fat diets and participated in routine levels of general physical activity (e.g., housework and walking). However, activities of AMI patients were restricted following admission. All participants (or their direct relatives) gave written informed consent, and the First Affiliated Hospital of Harbin Medical University and the Fourth Affiliated Hospital of Harbin Medical University approved all study protocols.

Sample Collection and DNA Extraction

Fresh fecal samples (each 2–5 g) were collected from all the participants under the hospital diet, then transferred into sterile collecting pipes and frozen at −80°C immediately. The associated clinical data were collected simultaneously. The bacterial DNA was extracted from the fecal samples using the TIANamp Bacteria DNA kit (Tiangen, Beijing, China) according to the manufacturer's instructions.

16S rRNA Gene V3–V4 Region Sequencing

DNA extracted from each sample was used as a template, and the V3–V4 region of the 16S rRNA gene was amplified using PCR. PCR amplification, sequencing of the PCR amplicons and quality control of raw data were performed. The purified products were mixed in equal proportions for sequencing.

Sequencing Data Analysis

First, we evaluated the quality of sequencing data using the Fast-QC software[1]. Second, clean Data were obtained for subsequent analysis after removing the Chimera Sequence using QIIME2[2]. Third, Operational taxonomic units (OTUs) were delineated at the cutoff of 97% also using QIIME2, and then the sequencing results were compared and analyzed to obtain the family and genus annotations of OTUs based on the Silva database[3]. Fourth, α- and β-diversity analyses were performed using

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

[2] http://qiime.org/

[3] https://www.arb-silva.de/

QIIME2. Shannon–Wiener diversity index, Simpson diversity index, the observed OTUs, PD (phylogenetic diversity)-whole-tree and Chao1 index were evaluated. A normalized OTU abundance table was used for the β-diversity analysis, including principal coordinate analysis (PCoA) based on weighted UniFrac, and unweighted UniFrac distances. Next, Lefse analysis was performed to clarify the dominant bacteria. LEfSe is a software for discovering high-dimensional biomarkers and revealing genome characteristics. LEfSe uses linear discriminant analysis (LDA) (Yang et al., 2020) to estimate the impact of the abundance of each component (species) on the difference effect. Finally, the gene function of the sample was inferred based on the species composition obtained by sequencing, and the functional difference between different groups was analyzed using PICRUSt[4]. Subsequently, the Welch's *t*-test method of two groups was performed using the STAMP software to filter the parts with *P*-value > 0.05, and Heatmap Plot, PCA plot, and Extended error bar graphs were drawn to reveal significant differences in species abundance between different samples. Based on the data obtained by sequencing, we performed the differential taxonomy expression analysis using limma algorithm for screening, and the differential screening criteria are: LogFC > 0.585 or <–0.585, *P*-value < 0.05.

RESULTS

Baseline Characteristics

A summary of the baseline characteristics of all the participants is presented in **Table 1**. AMI patients were characterized as consisting of a greater number of males, worsened cardiac functions, larger left ventricular end diastolic diameter (LVEDD), increased serum Troponin I (TnI) and NT-pro B-type natriuretic peptide (NT-proBNP) levels, increased numbers of leukocytes, neutrophils, and monocytes, increased fasting blood glucose levels and an increased prevalence of comorbidity with hypertension. In addition, among the patients enrolled in this study, including 15 ST-elevation myocardiol infarction (STEMI) and 15 non-ST elevation myocardiol infarction (NSTEMI), 19 experienced left anterior descending coronary (LAD) stenosis as the infarction related artery (IRA) and 21 had two or more coronary artery stenosis.

TABLE 1 | Baseline characteristics of the study participants.

Variables	AMI patients(*n* = 30)	Healthy controls(n = 30)	*P*-value
Age, years	62.6(9.02)	60.0(9.64)	0.2915
Sex, male	18(60%)	10(33%)	0.0390
BMI(kg/m^2)	25.4(3.33)	24.9(3.08)	0.5685
STEMI(%)	15(50%)	–	–
IRA (LAD)	19(63%)	–	–
≥2 coronary stenosis	21(70%)	–	–
Syntax score	18.1(5.95)	–	–
NYHA class (I/II/III/IV)	13/7/5/5	–	–
Hypertension	22(73%)	11(37%)	0.0038
Diabetes	9(30%)	5(16.7%)	0.2291
Atrial fibrillation	2(6.7%)	0(0)	0.1555
Smoking	13(43.3%)	6(20%)	0.0533
Echocardiogaraphic parameters			
LVEDD, mm	48.6(5.09)	43.7(3.99)	<0.0001
LVEF, %	51.7(8.78)	63.2(4.65)	<0.0001
SV, ml	52.2(12.06)	58.9(11.94)	0.0383
E/e'	64(60-68)	60(58–65)	<0.0001
Laboratory parameters			
TnI, ng/dL	27.6(0.10–226.9)	0.012(0–0.048)	0.0120
NT-proBNP, pg/mL	2652.1(113–26259)	124.0(25–258)	<0.0001
Leukocyte, 10^9/L	8.6(2.73)	6.7(1.74)	0.0019
Neutrophils, 10^9/L	6.0(2.62)	4.0(1.30)	0.0006
Lymphocytes, 10^9/L	2.1(1.17)	2.1(0.65)	0.9859
Monocyte, 10^9/L	0.7(0.31)	0.4(0.14)	<0.0001
Hemoglobin, g/L	133.6(34.77)	140.8(17.04)	0.3140
BUN, mg/dl	6.2(2.42)	5.5(1.81)	0.1714
Serum creatinine, mg/dl	76.1(35.38)	67.4(18.35)	0.2415
Fast glucose	6.6(2.21)	5.2(1.30)	0.0061
Cholesterol	4.6(1.33)	5.0(0.90)	0.1339
Triglycerides	1.8(0.76)	1.9(1.66)	0.5977
HDL-C	1.0(0.59)	43(58.11)	0.3584
LDL-C	2.6(0.87)	2.9(0.80)	0.1643
Uric acid	340.5(95.17)	316.9(73.03)	0.2853

Results are presented as median (with standard error or upper and lower quartiles) or % where appropriate.

BMI, body mass index; STEMI, ST-elevation myocardiol infarction; IRA, Infarction related artery; LAD, left anterior descending coronary; NYHA, New York Heart Association; LVEDD, left ventricular end diastolic diameter; LVEF, Left ventricular ejection fraction; TnI, Troponin I; NT-proBNP, NT-pro B-type natriuretic peptide; HDL-C, high density lipoprotein-cholesterol; LDL-C, low density lipoprotein-cholesterol.

Species Classification

Figure 1 contains a summary of the overall distribution of relative abundance of the top 20 phyla in each fecal sample (**Figure 1A**), as well as those found within each of the two groups (**Figure 1B**). Sequencing analysis revealed that the gut microbiota of the two groups were mainly contained within four phyla, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, and *Verrucomicrobia* (**Figure 1B**). The phylum with the highest abundance of reads in AMI patients was *Firmicutes*, accounting for 63.8% in total, versus that of an abundance of 72.4% in the controls (**Figure 1B**). The second greatest abundance in AMI patients was the phylum *Bacteroidetes*, accounting for 19.5% in total as compared with 17.7% in the controls (**Figure 1B**). Compared with the healthy control group, there was a rising trend but no significance of *Firmicutes* to *Bactericides* ratio in the patients with AMI (**Figure 1C**). Overall, there was a greater abundance in AMI versus controls for bacteria belonging to the phyla *Actinobacteria* (1.5 vs. 0.9%), *Cyanobacteria* (0.4 vs. 0.0%), *Proteobacteria* (9.6 vs. 6.8%), and *Verrucomicrobia* (5.0 vs. 1.5%). While a greater abundance was observed in controls versus AMI patients for bacteria belonging to the phyla *Fusobacteria* (0.55 vs. 0.1%) and *Tenericutes* (0.1 vs. 0.0%).

At the genus level, the microflora of AMI patients was characterized by lower levels of *Faecalibacterium*, *Roseburia*, *Tyzzerella 3*, [*Eubacterium*] *ventriosum group*,

[4] https://picrust.github.io/picrust/index.html

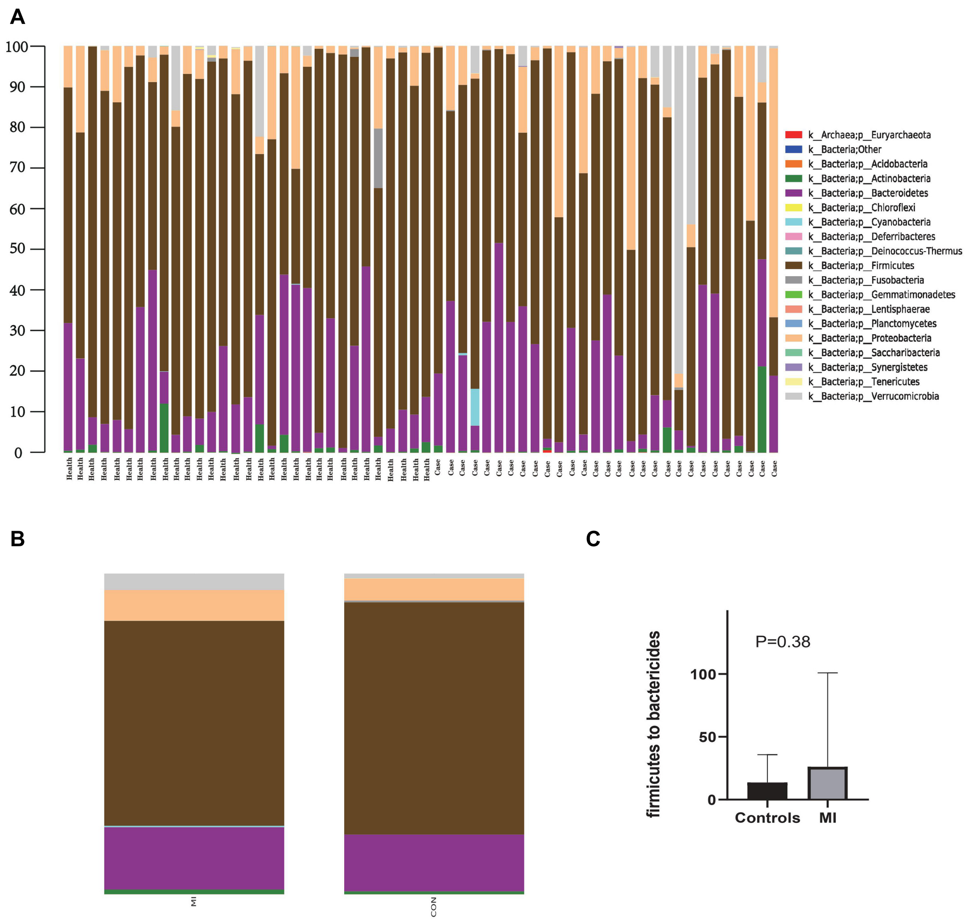


FIGURE 1 | The distribution of relative abundance of top 20 at the phylum level. **(A)** Shows the distribution of relative abundance of top 20 in each fecal sample. **(B)** Shows the distribution of relative abundance of top 20 in AMI group and the healthy control group. MI, AMI group; CON, the healthy control group. **(C)** Shows the *Firmicutes* to *Bactericides* ratio in AMI group and the healthy control group.

[Eubacterium] rectale group, *Ruminococcaceae NK4A214 group*, *Ruminococcaceae UCG-013*, *Ruminococcaceae UCG-014*, *Ruminococcus 1*, *Ruminococcus 2*, *Ruminococcaceae uncultured*, *Erysipelotrichaceae UCG-003*, *Megamonas*, *Fusobacterium*, and *Parasutterella*, and higher levels of *Bifidobacterium*, *Butyricimonas*, *Parabacteroides*, *Chloroplast; Other; Other; Other*, *Lysinibacillus*, *Lactobacillus*, *Christensenellaceae R-7 group*, *Subdoligranulum*, [*Eubacterium*] *coprostanoligenes group*, *Phascolarctobacterium*, *Megasphaera*, *Veillonella*, *Klebsiella*, and *Akkermansia* (**Supplementary Figure 1**).

Analysis of α and β Diversity Index

An α diversity analysis was performed and the resultant chao1, observed-outs, PD-whole-tree, Shannon–Wiener, and Simpson curves as based on the species annotation information obtained by sequencing analysis are presented in **Supplementary Figure 2**. The chao1 ($P = 0.0472$) and PD-whole-tree ($P = 0.0426$) indices were significantly decreased in the AMI versus control group (**Figure 2**). However, no statistically significant differences were obtained in the Shannon and Simpson indices. Taxonomic compositions of the metagenomic populations of gut microflora samples from AMI patients were compared with those from the healthy control group using the Principal Coordinate Analysis (PCoA). Differences in β-diversity as based on unweighted and weighted UniFrac values between the AMI and control group are shown in **Supplementary Figure 2**. The results of this analysis indicates that the fecal microbial structure of the AMI group differs from that of the healthy control group with regard to the presence of OTU.

Difference Expression Analysis Between the AMI and Control Groups

A differential taxonomy expression analysis was performed using limma algorithms. When focusing on differences at the genus level, our results revealed a remarkable difference with 50 generus in fecal microflora between the AMI and healthy control group. Among these changes, the increases in *Megasphaera*, *Butyricimonas*, *Acidaminococcus,* and *Desulfovibrio*, and decreases in *Tyzzerella 3*, *Dialister*, *[Eubacterium] ventriosum group*, *Pseudobutyrivibrio*, and

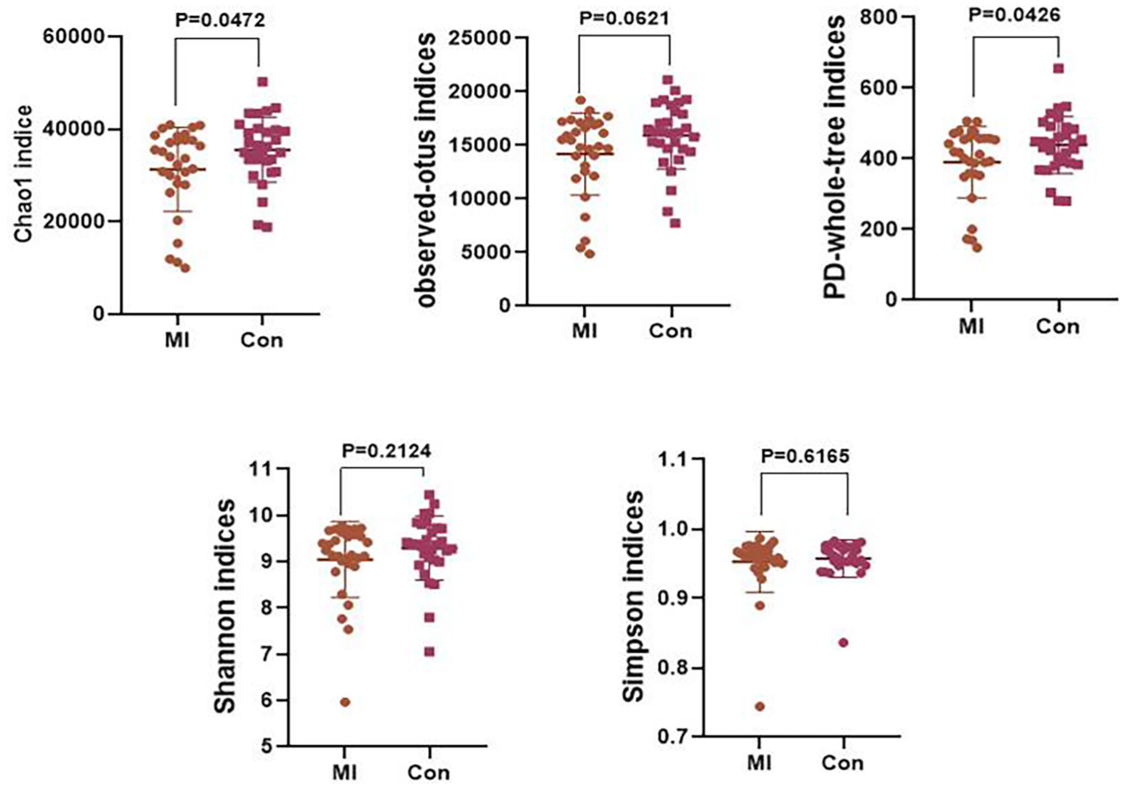


FIGURE 2 | The level of α-diversity indices of the gut microflora between the AMI group and the healthy control group. MI, AMI group; CON, the healthy control group.

Lachnospiraceae ND3007 group were the most notable features (**Figure 3**).

We also searched in the gutMDisoeder database (Cheng et al., 2020) whether the above gut microbes associated with AMI have the same pattern of changes as other diseases, or the opposite pattern of change with intervention measures, and the results were showed in **Tables 2**, **3**. The increase of genera *Megasphaera* were reported to be related to Parkinson's disease and metabolic syndrome, and the reduced abundance was observed by giving Vitamin D intervention. The increase of genera *Desulfovibrio* were reported to be associated with multiple sclerosis and gestational diabetes, and the reduced abundance was observed by giving *N*-acetylcysteine or dextran sulfate sodium intervention. The decrease of genera *Dialister* were reported to be related to asthma, thyroiditis, and type 1 diabetes mellitus, and the increased abundance was observed by giving polydextrose or soluble corn fiber intervention.

Differences Between the Subgroups in Patients With AMI

An analysis of the differences in the composition were performed among subgroups of AMI patients including, STEMI vs. NSTEMI, IRA-LAD vs. IRA-Non-LAD and Multiple (≥2 coronary stenosis) vs. Single coronary stenosis groups. Remarkable differences in fecal microflora were found among these subgroups.

Mean abundances of the genera *Pseudomonadales*, *Eubacterium-coprostanoligenes group,* and *Porphyromonadaceae* were greater in the STEMI vs. NSTEMI group, while a significantly greater abundance of the genera *Streptococcaceae*, *Lachnospiraceae NK4A136 group*, *Lactobacillus*, *Intestinibacter*, *Veillonella*, *Streptococcus*, *Peptostreptococcaceae*, *Erysipelatoclostridium*, *Veillonellaceae*, *Megasphaera*, *Lactobacillaceae*, *Peptoclostridium*, and *[Clostridium]innocuum group* were found in the NSTEMI vs. STEMI group (**Figures 4A,B**).

With regard to the IRA-LAD vs. IRA-Non-LAD subgroup, our results suggested a significantly greater abundance of the genera *[Eubacterium]ruminantium group*, *Comamonadaceae*, *Comamonas*, and the bacteria belonging to the order *MollicutesRF9*, as well as the bacteria belonging to the *Tenericutes* phyla in patients with LAD as the IRA vs. the IRA-Non-LAD subgroup. Whereas mean abundances of the genera *Veillonella*, *Bifidobacteriaceae*, *Bifidobacterium*, *Phocaeicola,* and others belonging to the *Lachnospiraceae* family, bacteria belonging to the *BacteroidalesIncertaeSedis* family and bacteria belonging to the *Actinobacteria* class were greater in the IRA-Non-LAD vs. IRA-LAD group (**Figures 4C,D**).

When comparing the Multiple (≥2 coronary stenosis) vs. Single coronary stenosis groups, a significantly greater

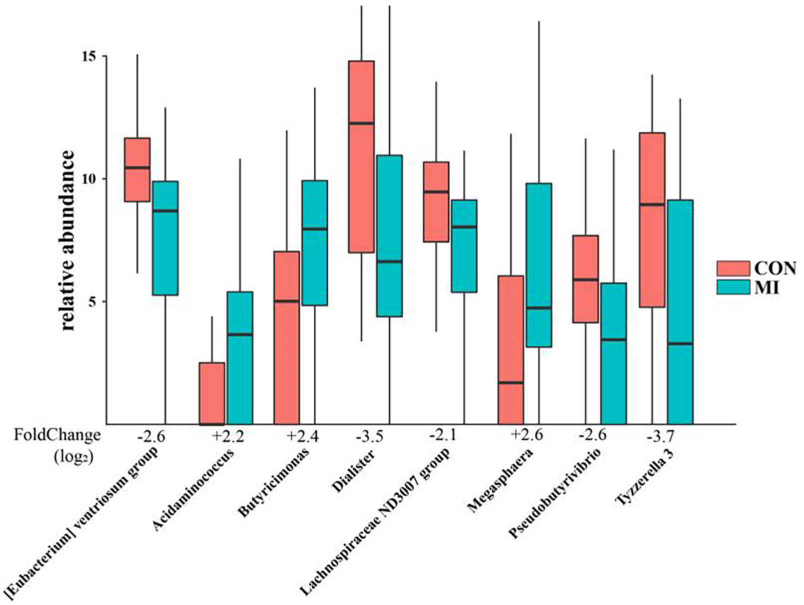


FIGURE 3 | The remarkable results of differential taxonomy expression analysis using limma algorithm between the AMI and healthy control group. MI, AMI group; CON, the healthy control group.

abundance of the genera *[Eubacterium]ruminantium group*, *Christensenellaceae*, *Christensenellaceae R-7 group*, *Collinsella,* and *Anaerotruncus*, as well as the bacteria belonging to *FamilyXI* were observed in the Single coronary stenosis group (**Figures 4E,F**).

Correlations Between the Clinical Characteristics and the AMI Microflora

Correlations were performed between the composition of gut microflora and significant clinical characteristics within AMI patients. Results of these analyses revealed that the genera *Bromus tectorum*, *Sphingomonas*, and *Candidatus Saccharimonas* were positively correlated with LVEDD, while the genera *Eisenbergiella* and *Ruminococcaceae NK4A214 group* were negatively correlated with LVEDD. The genera *Prevotellaceae UCG-001*, *Weissella*, *[Bacteroides] pectinophilus group*, *Veillonella*, *Rhizobium*, *Cronobacter*, *Lelliottia*, *Pseudocitrobacter,* and *Raoultella* were positively correlated with left ventricular ejection fraction (LVEF), while the genera *Parabacteroides* and *Sutterella* were negatively correlated with LVEF. The genus *Eubacterium* was positively correlated with levels of TnI, while the genus *Veillonella* was negatively correlated with TnI levels. The genus *[Clostridium] innocuum group* was positively correlated with levels of NT-pro-BNP, while the genera *Weissella* and *Veillonella* were negatively correlated with NT-pro-BNP levels. The genera *Alloprevotella*, *Prevotalla9*, *Chryseobacterium*, *Peptococcus*, *Romboutsia*, *Acidaminococcus*, *Phascolarctobacterium*, and *Achromobacter* were positively correlated with Syntax scores. The genera *Anaerofilum* and *Fastidiosipila* were positively correlated with leukocyte and neutrophil counts, *Olsenella* and *Cloacibacillus* were positively correlated with counts of neutrophils and the genera *Bacteroides*, *Phascolarctobacterium,* and *Bilophila* were negatively correlated with counts of monocytes. The genera *Eisenbergiella*, *Lolium perenne*, *Anaeroglobus,* and *Akkermansia* were positively correlated with fasting serum glucose levels, while *Lachnospiraceae NK4A136 group*, *Prevotella 2*, *Lachnospira*, *Tyzzerella 4,* and *Ruminococcaceae UCG-003* were negatively correlated with these fasting glucose levels (**Figure 5**).

Predictive Functional Analysis

PICRUst, as based on closed-reference OTU, was applied to predict abundances of the functional category COG orthologs (COs) and KEGG orthologs (KOs). Some of these COs and KOs demonstrated significantly different abundances in fecal microbiomes between the AMI and healthy control group (P < 0.05; **Figure 6**). Results from the COG database indicated that, inorganic ion transport and metabolism functioning, intracellular trafficking, secretion/vesicular transport, secondary metabolite biosynthesis, transport/catabolism and RNA processing/modification were all significantly increased in the AMI group. In contrast, defense mechanisms and cell cycle control functions, cell division and chromosome partitioning were significantly increased in the healthy control group (P < 0.05 for both groups; **Figures 6A,B**). However, at the level of KEGG pathways, we found significant increases in the AMI versus control group for the following processes: functioning of lipopolysaccharide biosynthesis proteins, membrane and intracellular structural molecules, biosynthesis of ubiquinone and other terpenoid-quinones, bacterial secretion, inorganic ion transport/metabolism, ionic pore channels, lipoic acid metabolism, tyrosine metabolism, the ubiquitin system, drug metabolism of cytochrome P450, aminobenzoate

TABLE 2 | The gut microbes associated with AMI have the same pattern of change as other diseases searched in the gutMDisoeder database.

Gut microbe	Alteration	Disorder
Megasphaera	Up	Parkinson's disease
Megasphaera	Up	Metabolic syndrome
Butyricimonas	Up	Infectious diarrhea
Acidaminococcus	Up	Parkinson's disease
Acidaminococcus	Up	Idiopathic calcium stone
Acidaminococcus	Up	Colorectal cancer
Acidaminococcus	Up	Inflammatory bowel disease
Desulfovibrio	Up	Microscopic colitis
Desulfovibrio	Up	Multiple sclerosis
Desulfovibrio	Up	Gestational diabetes
Desulfovibrio	Up	Familial adenomatous polyposis
Desulfovibrio	Up	Inflammatory bowel disease
Desulfovibrio	Up	Hepatic encephalopathy
Desulfovibrio	Up	Infectious diarrhea
Desulfovibrio	Up	Human immunodeficiency virus infectious disease
Dialister	Down	Asthma
Dialister	Down	Thyroiditis
Dialister	Down	Familial Mediterranean fever
Dialister	Down	Type 1 diabetes mellitus
Dialister	Down	Henoch-Schoenlein purpura
Dialister	Down	Spinal cord injury
Pseudobutyrivibrio	Down	Acute myeloid leukemia
Pseudobutyrivibrio	Down	Spinal cord injury
Lachnospiraceae ND3007 group	Down	Idiopathic calcium stone

TABLE 3 | The gut microbes associated with AMI have the opposite pattern of change with interventions searched in the gutMDisoeder database.

Intervention	Alteration	Gut microbe
Vitamin D	Down	Megasphaera
Bifico	Down	Desulfovibrio
JinQi Jiangtang	Down	Desulfovibrio
Perilla oil	Down	Desulfovibrio
N-acetylcysteine	Down	Desulfovibrio
Dextran sulfate sodium	Down	Desulfovibrio
Polydextrose	Up	Dialister
Soluble corn fiber	Up	Dialister

degradation, the citrate (TCA) cycle, tryptophan metabolism, geraniol degradation, protein folding and associated processing, amino acid metabolism, inositol phosphate metabolism, glutathione metabolism, limonene and pinene degradation, lipopolysaccharide biosynthesis, unsaturated fatty acid biosynthesis, valine, leucine and isoleucine degradation, biosynthesis/biodegradation of secondary metabolites, and fatty acid metabolism ($P < 0.05$; **Figures 6C,D**).

DISCUSSION

In the current study, fecal samples from 30 AMI patients and 30 healthy controls were collected to identify the composition and alterations in gut microbiota between these two groups as determined using bacterial 16S rRNA gene sequencing. Our results demonstrated a number of notable differences in gut microbial composition between these AMI patients and healthy controls. The composition of gut microflora was significantly correlated with clinical characteristics in AMI patients for such parameters as LVEDD, LVEF, serum TnI and NT-proBNP, Syntax scores, counts of leukocytes, neutrophils and monocytes, and fasting glucose levels. Moreover, significant differences in abundances of fecal microbiomes between the AMI and control group were obtained for some COs and KOs. We also analyzed differences in expressions among various subgroups of AMI patients. From this analysis we provide the first evidence indicating that remarkable differences in fecal microflora are present between the STEMI vs. NSTEMI, IRA-LAD vs. IRA-Non-LAD, and the Multiple (≥2 coronary stenosis) vs. Single coronary stenosis subgroups.

The gut microbiota, as the "second genome" of humans, is affected by a host of genes. Although the host genotype plays a decisive role in the composition and structure of the gut microbiota, the effect of diet cannot be ignored. Mounting evidence suggests that diet represents one of the most important factors influencing the composition and structure of gut microbiota (Cotillard et al., 2013), with changes in diet having the capacity to exert beneficial or harmful effects upon the composition of gut microbiota of the host. For example, it has been reported that a high-fat diet can damage the intestinal microbial environment and lead to microbiome dysregulation by reducing the amount of available carbohydrates in the colon, as well as increasing the level of intestinal oxygen stress and its own secondary metabolites (Ge et al., 2020; Li et al., 2020). In our study, all participants were from the same region, experienced a normal/routine lifestyle and had similar nutritional patterns, including typical Chinese diets based on carbohydrates versus high-fat diets. Furthermore, all the participants, including the healthy controls, were subjected to the hospital diet to minimize potential confounding effects of dietary differences on the microflora. We found that significant differences were present between AMI patients and healthy controls with regard to the fecal microbiome, suggesting the existence of a link between gut microflora dysbacteriosis and AMI. At the phylum level, *Firmicutes* and *Bacteroidetes*, the two most abundant phyla inhabiting the intestinal tract, are closely associated with environmental conditions and can be either beneficial or problematic to human and animal health. In addition, *Bacteroidetes* were reported to be implicated in immune regulation including activation of inflammation and autoimmune diseases (Carr et al., 2002; Gibiino et al., 2018; Nadia and Ramana, 2020). Our results suggest that the abundance of *Firmicutes* is decreased, while *Bacteroidetes* are slightly elevated in AMI patients. These findings are consistent with results obtained in an animal model of isoproterenol-induced AMI (Sun et al., 2019), but differ from results as obtained from fecal samples of patients with coronary heart disease (Kelly et al., 2016; Cui et al., 2017; Wang et al., 2018). One possible explanation for these results is that AMI, as a type of coronary heart disease with abrupt exacerbation and high

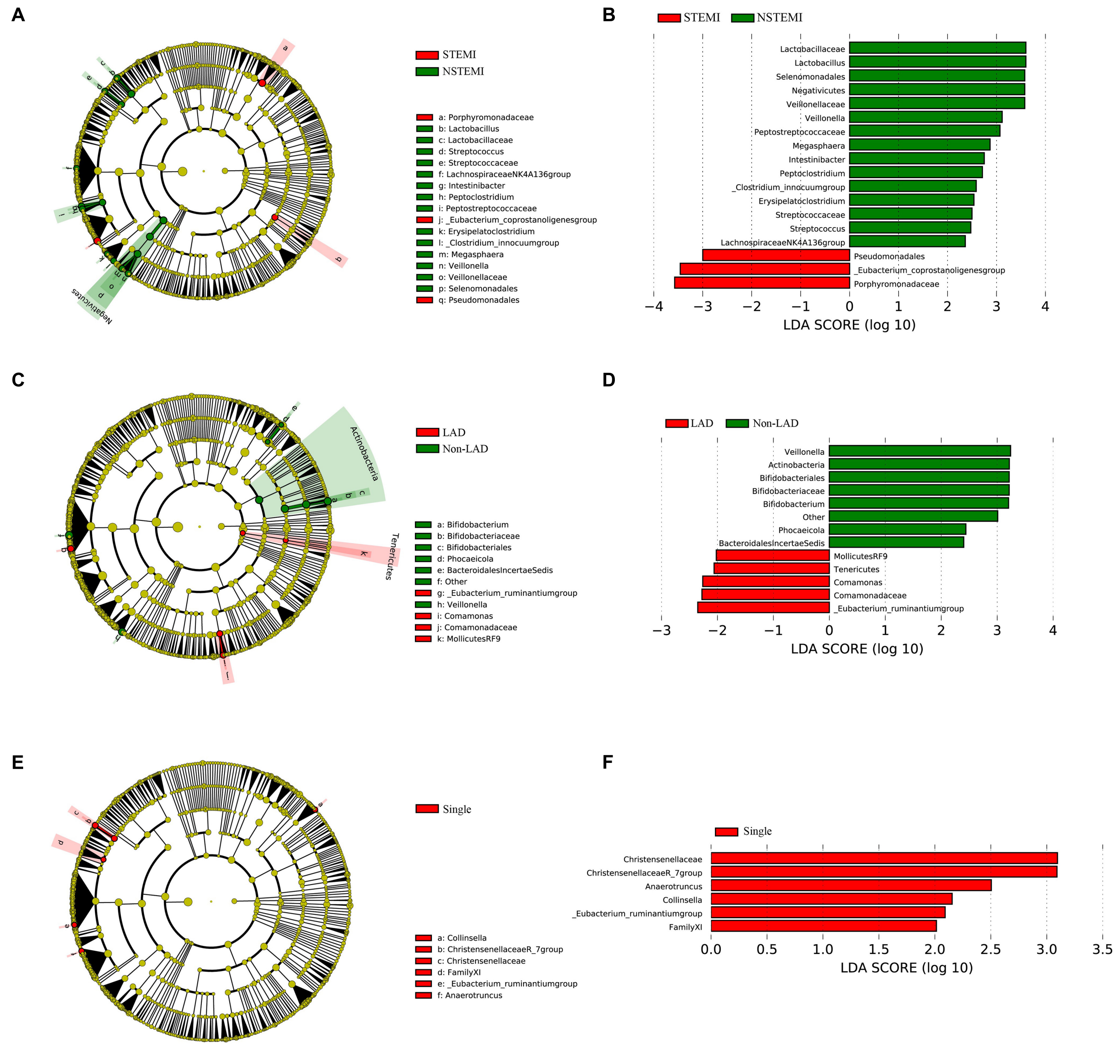


FIGURE 4 | The difference expression of gut microflora between the subgroups in patients with AMI. **(A,B)** Shows the different abundances of gut microflora between the STEMI group and the NSTEMI group; **(C,D)** shows the different abundances of gut microflora in the patients with LAD as the IRA compared to the IRA-Non-LADs; **(E,F)** shows the different abundances of gut microflora in the Single coronary stenosis group than in the Multiple (≥2 coronary stenosis) group. STEMI, ST-elevation myocardiol infarction; NSTEMI, Non ST-elevation myocardiol infarction; IRA, Infarction related artery; LAD, left anterior descending coronary.

mortality, has a unique pathophysiological process, including acute thrombosis, myocardial necrosis, inflammation, activation of the neuroendocrine system, and ventricular remodeling, may produce changes in gut microbiota. In addition, differences in gut environments may affect the abundance and composition of gut microbiota.

Microbiota diversity has emerged as a new biomarker of health (Shanahan, 2010; Cheng, 2019; Aponte et al., 2020; Ma et al., 2020). Loss of gut flora biodiversity is associated with various diseases, including active inflammatory bowel disease, childhood autism and recurrent *Clostridium difficile* associated diarrhea (Ott et al., 2004; Chang et al., 2008). In contrast, increased microbiota diversity is associated with enhanced health in the elderly (Claesson et al., 2012). Our current results show that the chao1 and PD-whole-tree indices of α-diversity were significantly decreased in the AMI group, revealing that the community richness of gut microbiota significantly decreased in the AMI patients. However, no statistically significant differences were obtained in other indices, including the Shannon index, results which may be attributable to the relatively small sample size.

Further differential taxonomy expression analyses using limma algorithms enabled us to focus on differences at the genus level. The AMI group was characterized by higher levels of *Megasphaera*, *Butyricimonas*, *Acidaminococcus,* and *Desulfovibrio*, and lower levels of *Tyzzerella 3*, *Dialister*, *[Eubacterium] ventriosum group*, *Pseudobutyrivibrio*, and

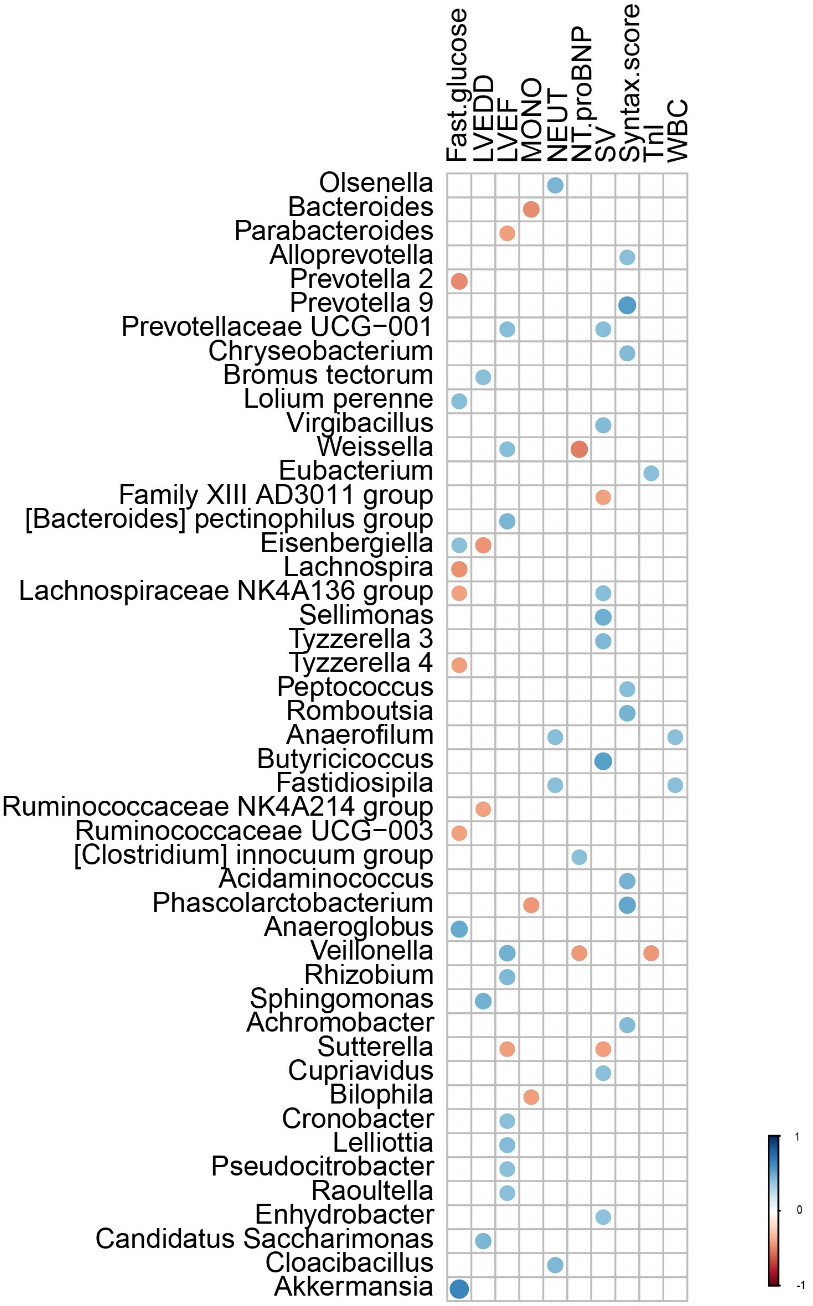


FIGURE 5 | The correlation between the gut microflora and the clinical characteristics with different significance in AMI patients. LVEDD, left ventricular end diastolic diameter; LVEF, left ventricular ejection fraction; TnI, Troponin I; NT-proBNP, NT-pro B-type natriuretic peptide; WBC, white blood cell; NEUT, neutrophils; LYMPH, lymphocytes; MONO, Monocyte.

Lachnospiraceae ND3007 group as compared with that observed in the healthy controls. *Megasphaera* belongs to the strictly anaerobic gram-negative cocci, which is involved in fermenting fructose and lactic acid with some short-chain fatty acids (SCFA) such as acetic acid and propionic acid main products as the main metabolites. *Butyricimonas*, which converts glucose into butyric and isobutyric acid, can also generate other types of SCFA such as acetic, propionic and succinic acid. These SCFA are an important source of energy for intestinal mucosal cells, contribute to the construction/repair of the intestinal mucosal barrier and resist oxidative stress (Sakamoto et al., 2014). *Desulfovibrio* can convert sulfates in food into sulfides, with hydrogen sulfide exerting a dual effect on gastrointestinal function through its capacity to either protect the gastrointestinal tract or participate in intestinal injury (Zhang-Sun et al., 2015). These bacterial genera have specific metabolic functions, and in individuals rich in these

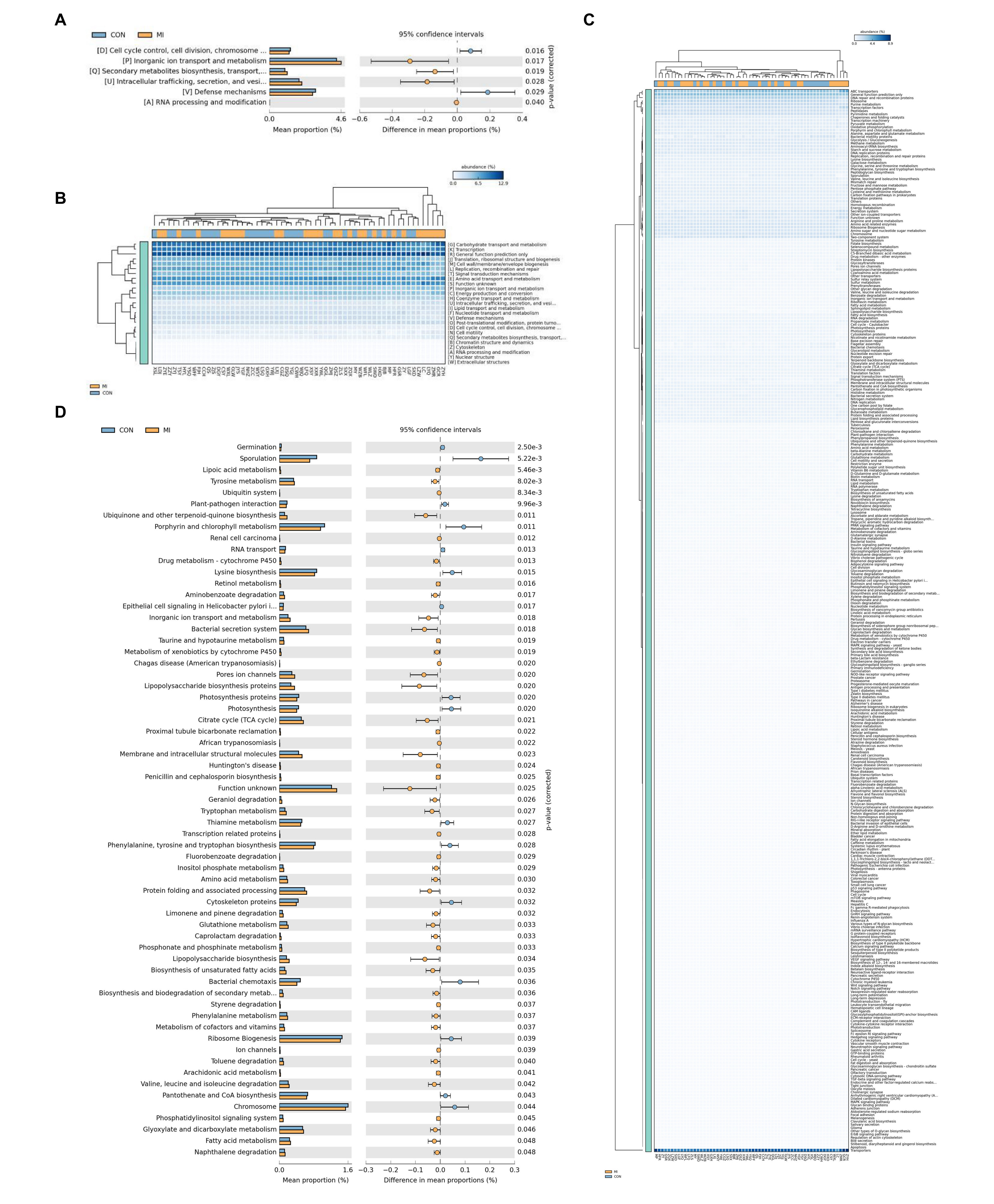


FIGURE 6 | The results of PICRUst based on closed-reference OTU to predict the abundances of functional categories COG orthologs (COs) and KEGG orthologs (KOs). **(A,B)** Shows the COs with significantly different abundances in the fecal microbiome between the AMI group and healthy control group; **(C,D)** shows the KOs with significantly different abundances in the fecal microbiome between the AMI group and healthy control group. MI, AMI group; CON, the healthy control group.

bacteria, lower levels of trimethylamine *N*-oxide (TMAO) are present, which is a major factor influencing cardiovascular diseases. *Acidaminococcus* is an anaerobic diplococcus that can use amino acids as their only source of energy for growth, which also belongs to the strictly anaerobic gram-negative cocci and produce acetic acid and butyric acid as metabolites. *Dialister* is one of the most representative types of intestinal flora associated with irritable bowel syndrome, and is believed to be correlated with dialister enrichment (Lopetuso et al., 2018). The main products are lactic acid, acetic acid and formic acid. *Tyzzerella* has been reported to be richly abundant in the patients with a high risk for cardiovascular diseases (Kelly et al., 2016; Xu et al., 2019), with the products of glucose fermentation include formic acid, lactic acid, acetic acid, ethanol, CO_2, and H_2. Only a few studies are available on the bacteria *[Eubacterium] ventriosum group* and *Lachnospiraceae ND3007 group*, accordingly, the physiological and pathological effects of most of these bacteria remain unclear. But it is certain that the main metabolites of these two genus are also SCFA. Even though the OTUs were assigned to the same genus, their functions may be distinct, as functions of bacteria are strain specific (Zhao, 2013). However, the common metabolites of the above-mentioned significantly different bacterial genera are mostly SCFA, which reveals that according to our results, the gut flora is most likely to affect the occurrence and development of AMI through the SCFA pathway. We also searched in the gutMDisoeder database and found that the pattern of changes in the gut microbes associated with AMI were the same as that of some other diseases such as Parkinson's disease, metabolic syndrome, multiple sclerosis, gestational diabetes and type 1 diabetes mellitus, or opposite to the pattern of changes in some interventions such as Vitamin D, *N*-acetylcysteine, dextran sulfate sodium, polydextrose, and soluble corn fiber. Most of these diseases are related to metabolism, obesity, immunity, and inflammation, and these factors also play an important role in the occurrence and development of AMI. Therefore, to a certain extent, it supports the hypothesis that change of gut flora participates in the pathophysiological process of AMI. Furthermore, some interventions such as Vitamin D and *N*-acetylcysteine might be useful for the treatment of AMI. These results will provide new direction for the role of intestinal flora in the pathophysiological process of AMI, as well as new targets for the treatment of AMI.

Our data also suggest that the composition of the gut microflora was significantly correlated with clinical characteristics of AMI patients, including LVEDD, LVEF, serum TnI and NT-proBNP, Syntax scores, WBC counts (neutrophils and monocytes), and fasting glucose levels. In specific, the correlations obtained indicated that gut microflora were associated with a greater incidence of LVEDD and lower incidence of LVEF suggesting that the gut microflora was involved with impaired cardiac function and left ventricular remodeling in AMI patients. And, the severity of AMI was characterized by serum levels of TnI and NT-proBNP. Furthermore, these indicators were significantly related to the prognosis of AMI patients, insinuating a role for gut microflora in the outcome of AMI patients. Among all gut microflora, the genera *Weissella* and *Veillonella* were positively correlated with LVEF and negatively correlated with levels of NT-pro-BNP indicating their role in cardiac functions. Similar to *Lactobacillus*, the genus *Weissella* was found to have a probiotic potential as a type of lactic acid bacteria (Anandharaj et al., 2015) with lactic acid and short-chain fatty acids as the metabolites, while the genus *Veillonella* has been reported to decompose lactic acid to produce propionic acid and promote metabolism (Scheiman et al., 2019), which also suggests that short-chain fatty acids play an important role in AMI. A sterile inflammatory environment is considered to be of paramount importance for AMI and ischemia/reperfusion injury development (Braunwald, 2015; Han et al., 2020), and is accompanied with elevated counts of leukocytes, neutrophils and monocytes. Like that as reported in other studies (Tang and Hazen, 2014; Yamashita et al., 2015; Amoroso et al., 2020; Wang et al., 2020), we found that gut microflora was related to immunity. The genera *Anaerofilum* and *Fastidiosipila* are both positively correlated with leukocyte and neutrophil counts, suggesting that they might be closely related to the sterile inflammatory conditions required for the pathophysiological process of AMI. In addition, gut microflora have been widely reported to be related with glucose and lipid metabolism (Qin et al., 2012; Karlsson et al., 2013; Fu et al., 2015), which is similar to our current results showing that gut microflora was related to fasting glucose levels.

To our knowledge, this is the first study that has examined differences in gut microflora among subgroups of AMI patients, especially the IRA-LAD vs. IRA-Non-LAD and Multiple (≥2 coronary stenosis) vs. Single coronary stenosis groups. Our results revealed that remarkable differences in fecal microflora were present between the STEMI vs. NSTEMI, IRA-LAD vs. IRA-Non-LAD and Multiple (≥2 coronary stenosis) vs. Single coronary stenosis groups. These findings not only indicate that gut microflora play an important role in the severity of AMI, but are also related to LAD occlusion and multiple coronary stenosis. Among the remarkable differences observed in fecal microflora of the subgroups, we found that nearly all of the bacterial genera belonged to the *Firmicutes* phyla in the NSTEMI and Single coronary stenosis groups, while most of the bacterial genera belonged to the *Proteobacteria* phyla in the STEMI and IRA-LAD groups. We were not able to further identify any direct correlations or mechanisms. In this way, the abundance of specific fecal microflora may possess the potential for prediction of pathophysiological and clinical characteristics of AMI.

Based on closed-reference OUT, PICRUSt was applied for predictive functional analysis. Several functional pathways, including inorganic ion transport and metabolism, secondary metabolite biosynthesis, transport, catabolism, protein folding and associated processing, amino acid metabolism, inositol phosphate metabolism and the Citrate (TCA) cycle have been identified. These functional pathways play an important role in such pathophysiological processes of AMI including myocardial necrosis, activation of acute inflammation, reperfusion injury and myocardial post-infarction repair. Therefore, these pathways can serve as a means to further predict the gut microflora that may contribute to AMI development, and it is probable due to its metabolite SCFA according to our results. SCFAs not only have the function of oxidative energy supply, but also have important functions such as maintaining water and electrolyte balance, anti-inflammatory, regulating immunity, regulating oxidative stress, anti-tumor and regulating gene expression. As the gut microbial ecosystem, which is considered as the largest endocrine organ of the body, can produce a variety of biologically active compounds that can be transported through the circulation and distributed to the distant parts of the host body, a plethora of basic biological and pathophysiological processes can impact the host (Tremaroli and Backhed, 2012). Therefore, long-term follow-up functional studies are urgently needed to reveal the specific bacteria that may contribute to the processes of AMI progression.

There are limitations with this current study. The relatively small sample sizes and lack of age/sex matched subjects for the AMI and control groups is a factor warranting consideration. The samples were only collected at a single time point, which precludes any assessments as to whether microbiota may fluctuate in patients with AMI during their treatment. Finally, multiple omics data will be required to further clarify the correlations between gut microbiota and AMI, as well as to establish the mechanisms through which gut microbiota affect the pathophysiological processes of AMI. In fact, it may be that multi-factorial processes are involved, which remains a subject for further investigation. Such determinations will likely need to be performed in animal models.

CONCLUSION

In conclusion, the present study suggested that the composition and the diversity of gut microflora were different between the AMI patients and healthy controls. Some fecal microflora were also found to be closely related to AMI clinical characteristics, as well as the alterations in the gut microbial community in different subgroups of AMI patients. Moreover, our results predicted several functional pathways based on the fecal microfloral information from AMI patients, which may enhance our comprehension of AMI pathogenesis. Overall, the process of AMI progression is dynamic and complicated, and modulation of the gut microbiota composition may represent a promising diagnostic biomarker or therapeutic target. In conclusion, the present results reveal that the composition and the diversity of gut microflora markedly differ between AMI patients and healthy controls. Since the common metabolites of the significantly different bacterial genera are mostly short-chain fatty acids, the gut flora is most likely to affect the occurrence and development of AMI through the short-chain fatty acid pathway. Some fecal microflora were found to be positively correlated with AMI clinical characteristics and distinct alterations in the gut microbial community were present within different subgroups of AMI patients. Moreover, our results show that predictions of several functional pathways can be generated as based on the fecal microfloral data from AMI patients. Such information may enhance our comprehension of AMI pathogenesis. Overall, the process of AMI progression is dynamic and complicated, and modulation of the gut microbiota composition may represent a promising diagnostic biomarker or therapeutic target.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the SRA database repository accession number PRJNA733305.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the First Affiliated Hospital of Harbin Medical University and the Fourth Affiliated Hospital of Harbin Medical University approved all study protocols. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HJi, PC, and HJu conceived and designed the experiments. JX, GS, and WS analyzed the data. CQ drew the pictures. YH and ZG wrote this manuscript. All the authors read and approved the final manuscript.

FUNDING

This study was supported by the National Key Research and Development Program of China (2019YFC0118100), National Natural Science Foundation of China (81671760 and 81873910), Excellent youth project of the Fourth Affiliated Hospital of Harbin Medical University (No. HYDSYYXQN202007).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.680101/full#supplementary-material

Supplementary Figure 1 | The distribution of relative abundance at the genus level.

Supplementary Figure 2 | α and β Diversity Index of the gut microflora.

REFERENCES

Alemao, C. A., Budden, K. F., Gomez, H. M., Rehman, S. F., Marshall, J. E., Shukla, S. D., et al. (2020). Impact of diet and the bacterial microbiome on the mucous barrier and immune disorders. *Allergy* 76, 714–734. doi: 10.1111/all.14548

Amoroso, C., Perillo, F., Strati, F., Fantini, M. C., Caprioli, F., and Facciotti, F. (2020). The role of gut microbiota biomodulators on mucosal immunity and intestinal inflammation. *Cells* 9:1234. doi: 10.3390/cells9051234

Anandharaj, M., Sivasankari, B., Santhanakaruppu, R., Manimaran, M., Rani, R. P., and Sivakumar, S. (2015). Determining the probiotic potential of cholesterol-reducing *Lactobacillus* and Weissella strains isolated from gherkins (fermented cucumber) and south Indian fermented koozh. *Res. Microbiol.* 166, 428–439. doi: 10.1016/j.resmic.2015.03.002

Aponte, M., Murru, N., and Shoukat, M. (2020). Therapeutic, prophylactic, and functional use of probiotics: a current perspective. *Front. Microbiol.* 11:562048. doi: 10.3389/fmicb.2020.562048

Braunwald, E. (2015). The war against heart failure: the lancet lecture. *Lancet* 385, 812–824. doi: 10.1016/s0140-6736(14)61889-4

Carr, F. J., Chill, D., and Maida, N. (2002). The lactic acid bacteria: a literature survey. *Crit. Rev. Microbiol,* 28, 281–370.

Chang, J. Y., Antonopoulos, D. A., Kalra, A., Tonelli, A., Khalife, W. T., Schmidt, T. M., et al. (2008). Decreased diversity of the fecal microbiome in recurrent clostridium difficile-associated diarrhea. *J. Infect. Dis.* 197, 435–438.

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D554–D560.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488, 178–184.

Cotillard, A., Kennedy, S. P., Kong, L. C., Prifti, E., Pons, N., Le Chatelier, E., et al. (2013). Dietary intervention impact on gut microbial gene richness. *Nature* 500, 585–588. doi: 10.1038/nature12480

Cui, L., Zhao, T., Hu, H., Zhang, W., and Hua, X. (2017). Association study of gut flora in coronary heart disease through high-throughput sequencing. *Biomed. Res. Int.* 2017:3796359.

Emoto, T., Yamashita, T., Sasaki, N., Hirota, Y., Hayashi, T., So, A., et al. (2016). Analysis of gut microbiota in coronary artery disease patients: a possible link between gut microbiota and coronary artery disease. *J. Atheroscler. Thromb.* 23, 908–921. doi: 10.5551/jat.32672

Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A., et al. (2015). The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* 117, 817–824. doi: 10.1161/circresaha.115.306807

Ge, X., Wang, C., Chen, H., Liu, T., Chen, L., Huang, Y., et al. (2020). Luteolin cooperated with metformin hydrochloride alleviates lipid metabolism disorders and optimizes intestinal flora compositions of high-fat diet mice. *Food Funct.* 11, 10033–10046. doi: 10.1039/d0fo01840f

Gibiino, G., Lopetuso, L. R., Scaldaferri, F., Rizzatti, G., Binda, C., and Gasbarrini, A. (2018). Exploring bacteroidetes: metabolic key points and immunological tricks of our gut commensals. *Dig Liver Dis.* 50, 635–639. doi: 10.1016/j.dld.2018.03.016

Grigoryan, H., Schiffman, C., Gunter, M. J., Naccarati, A., Polidoro, S., Dagnino, S., et al. (2019). Cys34 adductomics links colorectal cancer with the gut microbiota and redox biology. *Cancer Res.* 79, 6024–6031. doi: 10.1158/0008-5472.can-19-1529

Han, Y., Sun, W., Ren, D., Zhang, J., He, Z., Fedorova, J., et al. (2020). SIRT1 agonism modulates cardiac NLRP3 inflammasome through pyruvate dehydrogenase during ischemia and reperfusion. *Redox Biol.* 34:101538. doi: 10.1016/j.redox.2020.101538

Heianza, Y., Ma, W., DiDonato, J. A., Sun, Q., Rimm, E. B., Hu, F. B., et al. (2020). Long-term changes in gut microbial metabolite trimethylamine N-Oxide and Coronary heart disease risk. *J. Am. Coll. Cardiol.* 75, 763–772. doi: 10.1016/j.jacc.2019.11.060

Ingham, A. C., Kielsen, K., Cilieborg, M. S., Lund, O., Holmes, S., Aarestrup, F. M., et al. (2019). Specific gut microbiome members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell transplantation. *Microbiome* 7:131.

Jie, Z., Xia, H., Zhong, S. L., Feng, Q., Li, S., Liang, S., et al. (2017). The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* 8:845.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kelly, T. N., Bazzano, L. A., Ajami, N. J., He, H., Zhao, J., Petrosino, J. F., et al. (2016). Gut microbiome associates with lifetime cardiovascular disease risk profile among bogalusa heart study participants. *Circ. Res.* 119, 956–964. doi: 10.1161/circresaha.116.309219

Khan, M. Z., Munir, M. B., Khan, M. U., Osman, M., Agrawal, P., Syed, M., et al. (2020). Trends, outcomes, and predictors of revascularization in cardiogenic shock. *Am. J. Cardiol.* 125, 328–335. doi: 10.1016/j.amjcard.2019.10.040

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* 19, 576–585. doi: 10.1038/nm.3145

Kolodziejczyk, A. A., Zheng, D., and Elinav, E. (2019). Diet-microbiota interactions and personalized nutrition. *Nat. Rev. Microbiol.* 17, 742–753. doi: 10.1038/s41579-019-0256-8

Koren, O., Spor, A., Felin, J., Fak, F., Stombaugh, J., Tremaroli, V., et al. (2011). Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl 1), 4592–4598. doi: 10.1073/pnas.1011383107

Lam, V., Su, J., Hsu, A., Gross, G. J., Salzman, N. H., and Baker, J. E. (2016). Intestinal microbial metabolites are linked to severity of myocardial infarction in rats. *PLoS One* 11:e0160840. doi: 10.1371/journal.pone.0160840

Lam, V., Su, J., Koprowski, S., Hsu, A., Tweddell, J. S., Rafiee, P., et al. (2012). Intestinal microbiota determine severity of myocardial infarction in rats. *FASEB J.* 26, 1727–1735. doi: 10.1096/fj.11-197921

Li, X., Shi, W., Xiong, Q., Hu, Y., Qin, X., Wan, G., et al. (2020). Leptin improves intestinal flora dysfunction in mice with high-fat diet-induced obesity. *J. Int. Med. Res.* 48:300060520920062.

Liu, F., Fan, C., Zhang, L., Li, Y., Hou, H., Ma, Y., et al. (2020). Alterations of gut microbiome in tibetan patients with coronary heart disease. *Front. Cell Infect. Microbiol.* 10:373. doi: 10.3389/fcimb.2020.00373

Liu, Z., Li, J., Liu, H., Tang, Y., Zhan, Q., Lai, W., et al. (2019). The intestinal microbiota associated with cardiac valve calcification differs from that of coronary artery disease. *Atherosclerosis* 284, 121–128. doi: 10.1016/j.atherosclerosis.2018.11.038

Lopetuso, L. R., Petito, V., Graziani, C., Schiavoni, E., Paroni Sterbini, F., Poscia, A., et al. (2018). Gut microbiota in health, diverticular disease, irritable bowel syndrome, and inflammatory bowel diseases: time for microbial marker of gastrointestinal disorders. *Dig. Dis.* 36, 56–65. doi: 10.1159/000477205

Ma, X., Thakar, S. B., Zhang, H., Yu, Z., Meng, L., and Yue, J. (2020). Bioinformatics analysis of the rhizosphere microbiota of dangshan su pear in different soil types. *Curr. Bioinform.* 15, 503–514. doi: 10.2174/1574893615666200129104523

Marzullo, P., Di Renzo, L., Pugliese, G., De Siena, M., Barrea, L., Muscogiuri, G., et al. (2020). From obesity through gut microbiota to cardiovascular diseases: a dangerous journey. *Int. J. Obes. Suppl.* 10, 35–49. doi: 10.1038/s41367-020-0017-1

McCafferty, K., Byrne, C., and Yaqoob, M. (2012). Intestinal microbiota determine severity of myocardial infarction in rats. *FASEB J.* 26:4388. author reply 4388-9, doi: 10.1096/fj.12-1102ltr

Mendis, S., Thygesen, K., Kuulasmaa, K., Giampaoli, S., Mahonen, M., Ngu Blackett, K., et al. (2011). World Health Organization definition of myocardial infarction: 2008-09 revision. *Int. J. Epidemiol.* 40, 139–146. doi: 10.1093/ije/dyq165

Moraru, L., Moldovanu, S., Culea-Florescu, A.-L., Bibicu, D., Dey, N., Ashour, A. S., et al. (2019). Texture spectrum coupled with entropy and homogeneity image features for myocardium muscle characterization. *Curr. Bioinform.* 14, 295–304. doi: 10.2174/1574893614666181220095343

Morel, L., Domingues, O., Zimmer, J., and Michel, T. (2020). Revisiting the role of neurotrophic factors in inflammation. *Cells* 9:865. doi: 10.3390/cells9040865

Nadia, and Ramana, J. (2020). The human oncobiome database: a database of cancer microbiome datasets. *Curr. Bioinform.* 15, 472–477. doi: 10.2174/1574893614666190902152727

Nwokocha, C., Palacios, J., Simirgiotis, M. J., Thomas, J., Nwokocha, M., Young, L., et al. (2017). Aqueous extract from leaf of *Artocarpus altilis* provides cardio-protection from isoproterenol induced myocardial damage in rats: negative chronotropic and inotropic effects. *J. Ethnopharmacol.* 203, 163–170. doi: 10.1016/j.jep.2017.03.037

Ott, S. J., Musfeldt, M., Wenderoth, D. F., Hampe, J., Brant, O., Folsch, U. R., et al. (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut* 53, 685–693. doi: 10.1136/gut.2003.025403

Qi, C., Wang, P., Fu, T., Lu, M., Cai, Y., Chen, X., et al. (2021). A comprehensive review for gut microbes: technologies, interventions, metabolites and diseases. *Brief. Funct. Genomics* 20, 42–60. doi: 10.1093/bfgp/elaa029

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60.

Sakamoto, M., Tanaka, Y., Benno, Y., and Ohkuma, M. (2014). *Butyricimonas faecihominis* sp. nov. and *Butyricimonas paravirosa* sp. nov., isolated from human faeces, and emended description of the genus *Butyricimonas*. *Int. J. Syst. Evol. Microbiol.* 64, 2992–2997. doi: 10.1099/ijs.0.065318-0

Scheiman, J., Luber, J. M., Chavkin, T. A., MacDonald, T., Tung, A., Pham, L. D., et al. (2019). Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. *Nat. Med.* 25, 1104–1109. doi: 10.1038/s41591-019-0485-4

Shanahan, F. (2010). Probiotics in perspective. *Gastroenterology* 139, 1808–1812. doi: 10.1053/j.gastro.2010.10.025

Sonnenburg, J. L., and Backhed, F. (2016). Diet-microbiota interactions as moderators of human metabolism. *Nature* 535, 56–64. doi: 10.1038/nature18846

Sun, L., Jia, H., Li, J., Yu, M., Yang, Y., Tian, D., et al. (2019). Cecal Gut microbiota and metabolites might contribute to the severity of acute myocardial ischemia by impacting the intestinal permeability. Oxidative Stress, and Energy Metabolism. *Front. Microbiol.* 10:1745. doi: 10.3389/fmicb.2019.01745

Tang, W. H., and Hazen, S. L. (2014). The contributory role of gut microbiota in cardiovascular disease. *J. Clin. Invest.* 124, 4204–4211. doi: 10.1172/jci72331

Tang, W. H., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., et al. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* 368, 1575–1584. doi: 10.1056/nejmoa1109400

Thygesen, K., Alpert, J. S., Jaffe, A. S., Simoons, M. L., Chaitman, B. R., and White, H. D. (2012). Third universal definition of myocardial infarction. *Circulation* 126, 2020–2035.

Tremaroli, V., and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242–249. doi: 10.1038/nature11552

Vieira, S. M., Pagovich, O. E., and Kriegel, M. A. (2014). Diet, microbiota and autoimmune diseases. *Lupus* 23, 518–526. doi: 10.1177/0961203313501401

Wang, H., Liu, Y., Guan, H., and Fan, G.-L. (2020). The regulation of target genes by co-occupancy of transcription factors, c-Myc and mxi1 with max in the mouse cell Line. *Curr. Bioinform.* 15, 581–588. doi: 10.2174/1574893614666191106103633

Wang, W., Li, X., Yao, X., Cheng, X., and Zhu, Y. (2018). The characteristics analysis of intestinal microecology on cerebral infarction patients and its correlation with apolipoprotein E. *Medicine* 97:e12805. doi: 10.1097/md.0000000000012805

Xu, L., Huang, J., Zhang, Z., Qiu, J., Guo, Y., Zhao, H., et al. (2019). Bioinformatics study on serum triglyceride levels for analysis of a potential risk factor affecting blood pressure variability. *Curr. Bioinform.* 14, 376–385. doi: 10.2174/1574893614666190109152809

Yamashita, T., Kasahara, K., Emoto, T., Matsumoto, T., Mizoguchi, T., Kitano, N., et al. (2015). Intestinal immunity and gut microbiota as therapeutic targets for preventing atherosclerotic cardiovascular diseases. *Circ. J.* 79, 1882–1890. doi: 10.1253/circj.cj-15-0526

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yang, L., Gao, H., Wu, K., Zhang, H., Li, C., and Tang, L. (2020). Identification of cancerlectins by using cascade linear discriminant analysis and optimal g-gap tripeptide composition. *Curr. Bioinform.* 15, 528–537. doi: 10.2174/1574893614666190730103156

Zhang-Sun, W., Augusto, L. A., Zhao, L., and Caroff, M. (2015). Desulfovibrio desulfuricans isolates from the gut of a single individual: structural and biological lipid A characterization. *FEBS Lett.* 589, 165–171. doi: 10.1016/j.febslet.2014.11.042

Zhao, L. (2013). The gut microbiota and obesity: from correlation to causality. *Nat. Rev. Microbiol.* 11, 639–647. doi: 10.1038/nrmicro3089

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.

frontiers in Microbiology

MINI REVIEW
published: 19 July 2021
doi: 10.3389/fmicb.2021.683580

Applications of Raman Spectroscopy in Bacterial Infections: Principles, Advantages, and Shortcomings

Liang Wang[1]*, *Wei Liu*[2], *Jia-Wei Tang*[2], *Jun-Jiao Wang*[2], *Qing-Hua Liu*[3], *Peng-Bo Wen*[2], *Meng-Meng Wang*[4], *Ya-Cheng Pan*[5], *Bing Gu*[6]* *and Xiao Zhang*[2]*

[1] *Institute Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China,* [2] *Department of Bioinformatics, School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou, China,* [3] *State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa, China,* [4] *Jiangsu Key Laboratory of New Drug Research and Clinical Pharmacy, School of Pharmacy, Xuzhou Medical University, Xuzhou, China,* [5] *School of Life Sciences, Xuzhou Medical University, Xuzhou, China,* [6] *Laboratory Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China*

OPEN ACCESS

Edited by:
Amadeus Yeremia Pribowo,
Indonesia International Institute
for Life-Sciences (i3L), Indonesia

Reviewed by:
Piyush Baindara,
University of Missouri, United States
Miriam Cordovana,
Bruker Daltonik GmbH, Germany

***Correspondence:**
Xiao Zhang
changshui@hotmail.com
Bing Gu
gb20031129@163.com
Liang Wang
healthscience@foxmail.com

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

Received: *21 March 2021*
Accepted: *17 June 2021*
Published: *19 July 2021*

Citation:
Wang L, Liu W, Tang J-W, Wang J-J, Liu Q-H, Wen P-B, Wang M-M, Pan Y-C, Gu B and Zhang X (2021) Applications of Raman Spectroscopy in Bacterial Infections: Principles, Advantages, and Shortcomings.
Front. Microbiol. 12:683580.
doi: 10.3389/fmicb.2021.683580

Infectious diseases caused by bacterial pathogens are important public issues. In addition, due to the overuse of antibiotics, many multidrug-resistant bacterial pathogens have been widely encountered in clinical settings. Thus, the fast identification of bacteria pathogens and profiling of antibiotic resistance could greatly facilitate the precise treatment strategy of infectious diseases. So far, many conventional and molecular methods, both manual or automatized, have been developed for *in vitro* diagnostics, which have been proven to be accurate, reliable, and time efficient. Although Raman spectroscopy (RS) is an established technique in various fields such as geochemistry and material science, it is still considered as an emerging tool in research and diagnosis of infectious diseases. Based on current studies, it is too early to claim that RS may provide practical guidelines for microbiologists and clinicians because there is still a gap between basic research and clinical implementation. However, due to the promising prospects of label-free detection and noninvasive identification of bacterial infections and antibiotic resistance in several single steps, it is necessary to have an overview of the technique in terms of its strong points and shortcomings. Thus, in this review, we went through recent studies of RS in the field of infectious diseases, highlighting the application potentials of the technique and also current challenges that prevent its real-world applications.

Keywords: Raman spectroscopy, bacterial pathogen, machine learning, infectious disease, antibiotic resistance

INTRODUCTION

Infections caused by bacterial pathogens in clinical settings are commonly encountered, which is considered as the top 10 most common causes of death globally (Abayasekara et al., 2017). Acute bacterial infections could be serious or even deadly, especially when bacteria enter into bloodstream or cross the blood–brain barrier (van Sorge and Doran, 2012). In addition, antibiotic resistance plays important roles in bacterial pathogenicity during host infection. Thus, rapid detection of bacterial infection and profiling of drug resistance are crucial in guiding effective treatments of infectious diseases (Burnham et al., 2017). Conventional methods for bacterial diagnosis, such as medium culture, biochemical reactions, and serological tests are well-established techniques with high reliability and accuracy. However, some of these techniques are laborious, costly, and time consuming, which also have comparatively steep learning curves for real-world use (Franco-Duarte et al., 2019). Thus, new diagnostic methods have been developed for rapid and minimally invasive detection of bacterial pathogens in order to meet clinical requirements or investigate infectious disease outbreaks (Fournier et al., 2014), such as polymerase chain reaction (PCR), enzyme-linked immunosorbent assay (ELISA), high-throughput next-generation sequencing (NGS), and chemical analysis methods like mass spectrometry (MS). In recent years, Raman spectroscopy (RS) is gaining more and more attentions in research fields and in clinical settings due to advancements in instrumentation and data-handling techniques (Wang et al., 2016). As an easy-to-learn, low-cost, and label-free chemical analysis technique, RS has both great potentials and huge challenges in clinical pathogen analysis (Sil et al., 2020), which drives researchers to work hard in the field to bridge the gap between experimental setup and clinical implementation. In this review, we focus on the principles, advantages, and shortcomings of the RS technology in a concise manner, highlighting the application potentials of the technique and also current challenges that prevent its real-world applications.

CONVENTIONAL AND MOLECULAR METHODS

Conventionally, the detection of bacterial pathogens in clinical infection relies on methods like medium culture (e.g., colony size, color, and shape), microscopy (e.g., Gram stain), biochemical analysis (catalase activity, oxidase activity, and urease activity, etc.), and serological tests (e.g., latex agglutination tests; Váradi et al., 2017). The presence of antibiotic resistance adds more complexity during the identification of bacterial infections. Classical methods for antibiotic susceptibility testing (AST) include but not limited to disk diffusion, Epsilometer test (*E*-test), and microdilution, which also require medium culture (Khan et al., 2019). However, only a small number of bacteria could be successfully cultured due to the fastidious growth requirement, which makes accurate diagnosis of bacterial infection a challenge.

The development of molecular diagnostic methods greatly improves bacterial identification and antibiotics profiling, which mainly relies on the analysis of genomic markers corresponding to nucleic acid sequences (Váradi et al., 2017). For example, PCR is a fast and reliable molecular method for the identification of bacterial infections, which directly detects bacterial pathogens by genetic materials and requires primers for the amplification process (Barghouthi, 2011). One of the advantages of PCR is its capacity in recognizing bacterial infection at early stage when no sufficient antibodies against the pathogens are produced (Kubina and Dziedzic, 2020). However, once the pathogens are cleared from the immune system or become dormant, no genetic materials could be detected. Thus, PCR is better to be used during the acute infection stage and cannot be used for retrospective analysis. In recent years, universal primers with the capacity of identifying highly conservative regions of genes like 16s rDNA have also been widely used to find previously unrecognized or uncultured organisms from infected host tissues, leading to the characterization of microbial diversity within a sample, which is also known as metagenomics (Abayasekara et al., 2017).

Enzyme-linked immunosorbent assay is a type of immunosorbent assay that can be used for bacterial identification through detecting the presence of antigens or antibodies in blood sera. In the food industry, ELISA is one of the most commonly used immunological methods for foodborne pathogen detections (Law et al., 2015). Recently, some comparative studies indicated that ELISA had great potential in clinical applications due to its superiority to conventional methods in the diagnosis of bacterial infections (Xu et al., 2020). However, antibody levels in the early stage of post-infection may not be reliably detectable. Currently, ELISA has not been used in routine bacterial diagnostics, which may be due to its limitations such as high costs, poor reproducibility, high false-positive rates, and antibody instability (Sakamoto et al., 2017). One advantage of the serological testing via ELISA method, when compared with other methods based on genetic materials, is that it is able to study the epidemiology of diseases in different populations retrospectively due to the persistence of antibodies in the bloodstream after microbial infections (Lai et al., 2020).

As for the sequencing technology, it used to be difficult to access but is now affordable in microbial studies due to the fast development of instruments (Kwong et al., 2015). For example, NGS and long-read sequencing could provide high-resolution discrimination of bacterial pathogens at nucleic acids level, which could reliably distinguish closely related bacterial lineages and accurately track the outbreaks (Balloux et al., 2018; Logsdon et al., 2020). In addition, through comparative genome analysis, gain or loss of particular genes could be used to predict specific phenotypes such as stress resistance and pathogenicity (Stratakos et al., 2019), while genome-wide association study could reveal antibiotic resistance (Lees et al., 2020). As for microbial composition in a clinical sample from mouth, skin, or gut, metagenomic next-generation sequencing plays a pivotal role, which greatly facilitates the understanding of antimicrobial resistance, microbiome, human host gene expression, and oncology (Chiu and Miller, 2019). Although NGS provides an overview of bacterial species at genomic level with high accuracy, sequencing is still far away to be a routine method because of the high costs, labor intensity, complex sample preparation steps, and sophisticated data analysis procedures (Deurenberg et al., 2017). Currently, the application of NGS methods is mainly limited to laboratory experiments and epidemiological investigations while being rarely used for routine microbial identification or susceptibility testing in clinical laboratories (Deurenberg et al., 2017; Rossen et al., 2018).

CHEMICAL ANALYSIS METHODS

Mass Spectroscopy

In recent years, chemical analysis via precision instruments is getting more and more attention from both industrial, clinical, and academic fields, among which MS is an important analytical tool due to its high-throughput capacity, sensitivity, and specificity (Sauer and Kliem, 2010). Although several MS methods, together with software tools, have been developed, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is one of the most popular MS instruments used in clinical microbiology due to the rapid and accurate identification of an extensive range of bacteria (Hou et al., 2019). In specificity, MALDI-TOF MS is an inexpensive and straightforward method for bacterial classification and identification on genus, species, and, sometimes, subspecies level (Sauer and Kliem, 2010). In addition, databases containing MS spectra of known organisms provide much more convenience in the identification of species with similar phenotypic, genotypic, and biochemical properties (Singhal et al., 2015). However, there are also some limitations for MALDI-TOF MS. For example, it is difficult for MALDI-TOF MS to discriminate closely related bacterial species such as *Escherichia coli* and *Shigella*. It is also hard for MALDI-TOF MS to differentiate some antibiotic resistance phenotypes, such as methicillin-resistant and methicillin-sensitive *Staphylococcus aureus* (Florio et al., 2018).

Raman Spectroscopy

Raman spectroscopy is an emerging method for the identification of bacterial infections because it can act as a rapid, efficient, and minimally invasive tool to identify bacterial cells and antibiotic resistance, which also has the potentials in high-throughput and real-time applications in the field of clinical diagnostics (Strola et al., 2014). The basic principle of the Raman effect is that when the smallest unit of light passes through any medium, the light scattered by other molecules affects the frequency change, which means the Raman effect is caused by the vibration of molecules and thus can be explained by energy levels (Jones et al., 2019). From the perspective of quantum mechanics, the Raman effect is the inelastic collision that occurs when photons collide with molecules. If the molecule is at the ground state level at the beginning, and then when the excited light interacts with the molecule, the molecule will be excited to a high energy level or virtual state, and then, the molecules and electrons in the virtual state will transition to the excited state, generating scattered light. In this process, energy will be transferred to the molecule by the excited photon, while the photon loses its energy in this process. At this time, the molecule that transitions to the excited state gains energy.

There are some positions where the incident light frequency is at low level, and at these positions, the accepted scattered light is called Stokes Line, while, on the contrary, it is called anti-Stokes Line (Jones et al., 2019). When photons collide with molecules, the energy between them does not change after the collision, but the direction is changed, which is called Rayleigh scattering (Bumbrah and Sharma, 2016). Normally, RS has a strong fluorescence background that could disturb the original spectrum, which leads to compromised quality of bacterial identification, although it could be removed by techniques like polynomial baseline fitting (Wei et al., 2015). As for the application, RS produces a series of spectral signal lines when measuring a particular sample, in which Raman shift is the frequency difference between the Raman scattered light and the aforementioned Rayleigh scattered light (Cialla-May et al., 2019). Some specific molecules in biological samples will have characteristic peaks, and the concentration or amount of a certain molecule in the sample will affect the intensity of the molecule (**Figure 1**).

BACTERIAL IDENTIFICATION

Raman spectral features are generated by molecular vibrations in the sample, which makes RS a convenient tool for characterizing biological systems (Ashton et al., 2011). Due to its low-cost, label-free, and nondestructive features, RS has been widely investigated in terms of its potential applications in clinical studies. In addition, the sample preparation steps are simple, and the spectroscopic procedures can be completed within seconds, which makes it a promising method for detecting bacterial infection (Boardman et al., 2016). Through searching keywords RS and bacteria identification on the biomedical literature database, PubMed[1], it has been observed that there is a continuously growing number of RS-assisted bacterial detection studies. However, there is still a hug gap between basic research and practical application, which prevents RS from becoming the routine laboratory technique. For example, Raman effect is very weak, which leads to long measurement times; moreover, sample fluorescence introduces noisy signals into the spectrum, which makes the downstream analysis rather difficult (Wei et al., 2015). In addition, intense laser radiation can cause sample heating, leading to sample destruction and disrupted Raman spectrum. Thus, biological samples should be investigated via low-energy near-infrared wavelength for excitation, e.g., 785 or 830 nm, or in water solutions (Eberhardt et al., 2015).

During bacterial sample analysis, RS provides information on both chemical compositions and biomolecular structures, such as DNA, RNA, proteins, lipids, and carbohydrates, which is often referred as whole-organism fingerprint (Ashton et al., 2011). For example, Raman signal for C–H stretching vibration is at approximately 2,930 cm^{-1} and C–H deformation vibration at approximately 1,440 cm^{-1}, while the main signal for proteins is the amide I vibration at 1,665 cm^{-1} (Lorenz et al., 2017). In addition, RS has also been applied in single bacterial analysis and live bacterial studies, which could not only minimize the bacterial metabolic variability at the different phases but also facilitate the understanding of cellular dynamics (Strola et al., 2014; Smith et al., 2016). In terms of the differentiation of Gram-positive bacteria from Gram-negative bacteria, it was showed that some peaks at 540 and 1,380 cm^{-1} had significant differences

[1] https://pubmed.ncbi.nlm.nih.gov/

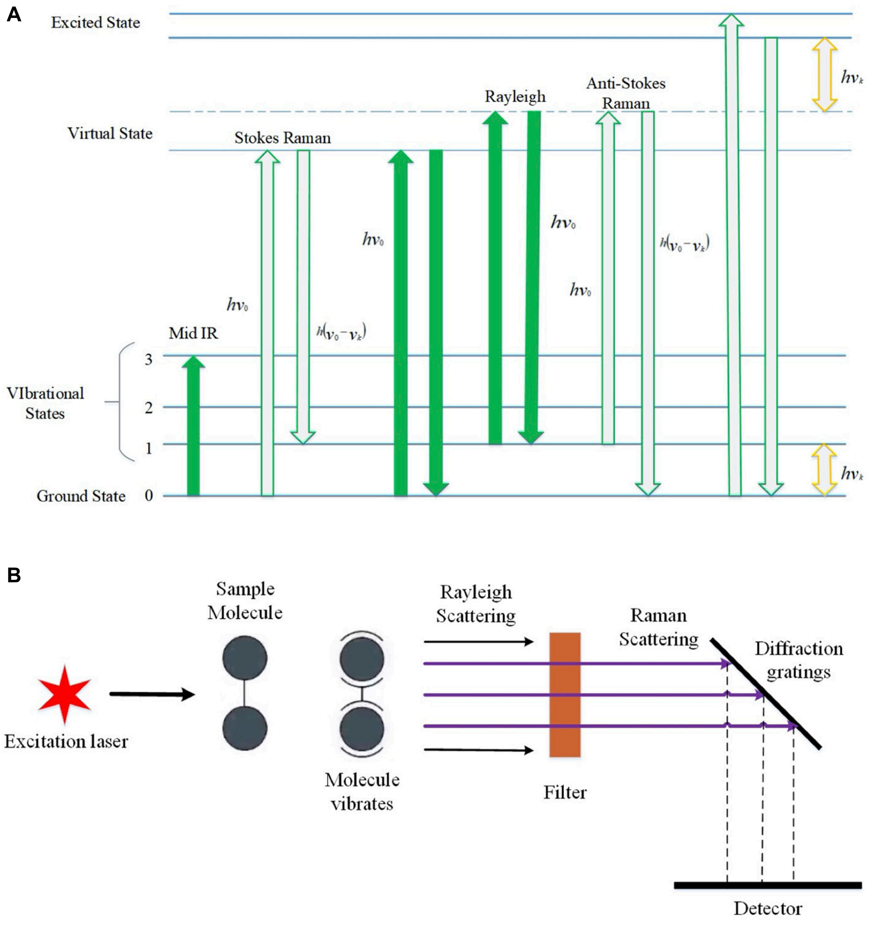


FIGURE 1 | Schematic illustration of the basic principles of Raman effects and the brief architecture of Raman spectroscopy. **(A)** Raman spectrum energy level diagram, which shows the transition process of infrared light irradiation, Stokes rays, anti-Stokes rays, Rayleigh scattering, and Raman scattering. hv_k, initial irradiation energy; E_0, ground state; E_1, vibration excited state; E_0+hv_0 and E_1+hv_0, excited virtual state. **(B)** Schematic diagram of Raman spectroscopy. After the incident light is irradiated, the molecules reach an excited state. The light of different frequencies during the scattering process is Raman scattering, which is reflected on the grating and captured by the detector.

for Gram-positive bacteria when compared with the Gram-negative bacteria, which was mainly attributed to the glycosidic bonds in *N*-acetyl glucosamine and *N*-acetyl muramic acid of peptidoglycan (de Siqueira E Oliveira et al., 2020).

So far, studies performing RS on clinical bacterial pathogens require culture in agar plates because of the low concentration of bacteria in clinical samples (Rebrošová et al., 2017; de Siqueira E Oliveira et al., 2020). Although culture-based RS could provide sufficient biomass during testing, hence higher signal–noise ratio, it is rather time consuming. There are also some attempts of RS applications on tissues in terms of infectious disease diagnosis *in situ*. Kloß et al. (2014) used RS and chemo-metrical evaluation to study the ascitic fluid directly for pathogen identification, which showed that 97.7% of the spectra from Gram-positive bacteria were correctly assigned on the genus level and 83.6% on the species level. In another study, Maquelin et al. (2003) used Raman spectra for rapidly identification of bacterial and fungal pathogens recovered from 115 blood cultures after 6- to 8-h culture in an automated blood culture system, according to which 109 samples contained bacteria while 6 contained yeasts (92.2% identification accuracy). Thus, RS possesses the potential in the identification of bacterial infections directly for clinical samples.

In some situations, clinical samples only contain trace amounts of bacterial cells. In order to improve the weak Raman signals in clinical samples such as blood and urine when bacterial amount is rather low, surface-enhanced Raman spectroscopy (SERS) can be applied, which could facilitate the development of culture-free identification of bacterial pathogens. In specificity, SERS is an enhanced RS through sample molecules interacting with surface plasmons of nanoscale structured metal surfaces, which often uses spherical nanoparticles made of silver or gold with diameters ranging from 20 to 100 nm (Krafft and Popp, 2014). For example, Tien et al. (2018) investigated bacterial pathogens in 108 urine samples sourced from of urinary tract infection patients; according to the study, 93 samples were detected with single bacterial species via SERS, while 97 samples were confirmed pathogen positive through medium culture, which makes the detection 95.87% accurate. Currently, although

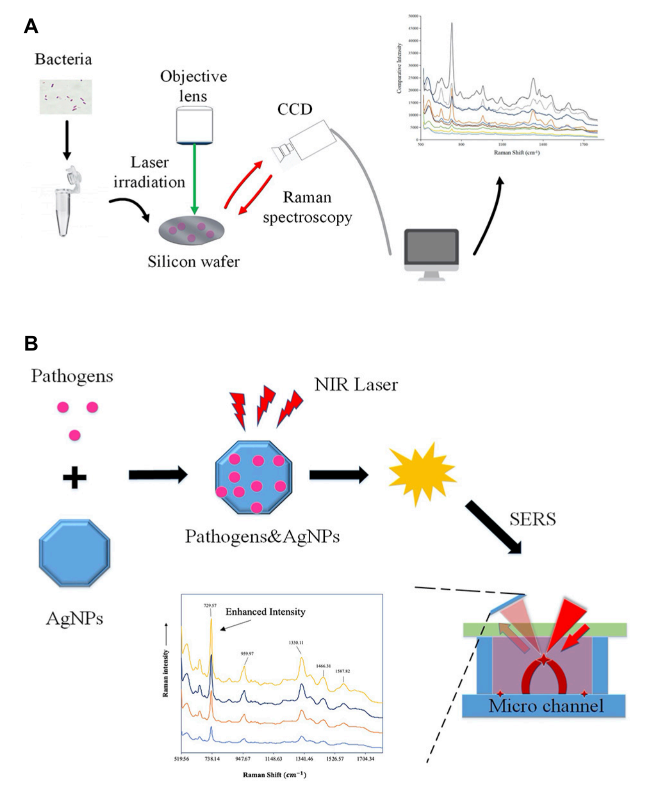


FIGURE 2 | Schematic illustration of the workflow of Raman spectroscopy and surface-enhanced Raman spectroscopy (SERS). **(A)** Workflow of Raman spectroscopy obtaining information on molecular structure via molecular vibrations and rotations for bacterial classification and antibiotic resistance profiling. **(B)** Procedures of the surface-enhanced Raman spectroscopy. Samples were measured while being adsorbed on the surface of colloidal metal nanoparticles such as silver (AgNPs), gold, or copper in order to improve signal intensity.

SERS is a highly promising analytical technique, it has not been used as a routine diagnostic method in the clinical laboratory yet, and there are many problems preventing its real-world application. One of the major limitations is the fabrication of suitable substrates with unique features in SERS-related detections, although tremendous effort has been invested into this area (Ouyang et al., 2017). Thus, developing new cost-effective and reproducible substrates for SERS would also greatly increase its sensitivity and accuracy, hence wider applications of the technique. Many studies have reported a variety of preparation procedures of nanoparticles for SERS (Solís et al., 2017; Demirtaş et al., 2020). However, this topic is rather large and is not a focus of the current review. For details, please refer to the recent review by Lane et al. (2015).

Except for bacterial infections, RS has also been applied for the identification of other microbial species, which shows great promise for the accurate diagnosis of parasites and viruses (Chen et al., 2016; Yeh et al., 2020; Donald et al., 2021). In particular, since the global outbreak of coronavirus 2019 (COVID-19), a variety of studies focus on the rapid detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) via RS. For example, Carlomagno et al. (2021) reported a Raman-based method for saliva analysis, which is able to differentiate healthy individuals from infected patients with accuracy, precision, sensitivity, and specificity of more than 95%. In addition, Yin et al. (2021) analyzed 177 serum samples (63 confirmed COVID-19 patients, 59 suspected cases, and 55 healthy individuals) via RS, together with 20 independent individuals for external validation. According to the study, accuracy between the COVID-19 and the healthy controls is 0.90, which also indicated that RS held the promise of being a safe and efficient technique for COVID-19 screening (Yin et al., 2021). For a schematic illustration of the workflow of Raman spectroscopy and surface-enhanced Raman spectroscopy (SERS), please refer to **Figure 2**.

ANTIBIOTIC RESISTANCE PROFILING

As for AST, it is an essential procedure in the clinical diagnosis of serious bacterial infection, while accurate and effective diagnosis of bacterial antibiotic resistance is a key for the treatment of bacterial infections (Wang et al., 2018). Although the typical procedure normally takes 3–4 days or even longer for fastidious bacteria on average to obtain the final AST results (Han et al., 2020), with MALDI-TOF MS-based approaches, i.e., for positive blood culture bottles, a result can be obtained after <24 h, in some cases also the same day (Verroken et al., 2014). Due to its simple operations, RS, especially SERS, has been used for testing antibiotic resistance phenotypes in many bacterial species, such as *E. coli* (Chang et al., 2019), *S. aureus* (Uysal Ciloglu et al., 2020), and *Pseudomonas aeruginosa* (Li et al., 2019). A variety of signatures have been observed in terms of bacterial antibiotic resistance and susceptibility, which could be used for rapidly identifying resistance to sublethal concentrations of antibiotics (Galvan and Yu, 2018; Han et al., 2020). In addition, a single study also reported that a portable Raman spectrometer with paper-based SERS could be used for screening tetracycline residues in milk with peak intensity ratios at 455 cm^{-1}/1,280 cm^{-1} and 874 cm^{-1}/1,397 cm^{-1}. Thus, RS could function as a potential tool for on-site monitoring of antibiotics (Marques et al., 2019). However, despite that SERS was investigated to detect antibiotic-resistant phenotypes in some studies, current datasets are small, limited, and often involving environmental settings. In addition, the ability of RS to detect resistance phenotypes is something different from antibiotic resistance testing, which does not rely on the presence of resistance markers but on the determination of minimum inhibitory concentration (Galvan and Yu, 2018). Thus, the generalization of these Raman signatures, biomarkers, or metabolites in predicting antibiotic resistance profiles should be further examined before applied in clinical settings.

COMPUTATIONAL ANALYSIS OF RAMAN SPECTRA

Due to the complexity of Raman spectra, statistics and machine learning algorithms, rather than traditional linear analysis, are normally involved in data processing procedures. So far, many machine learning methods have been introduced into Raman spectra analysis, such as artificial neural network, deep learning, and Monte Carlo estimation (Lu et al., 2012; Moawad et al., 2019; Lussier et al., 2020; Uysal Ciloglu et al., 2020). In the rapid characterization of *Staphylococcus*, Rebrošová et al. (2017) compared three machine learning methods, namely, linear discriminant analysis, one nearest neighbor, and support vector machine (SVM), all of which showed efficient identification of staphylococci using RS with high accuracy. Although machine learning often gives promising results during the analysis of Raman spectra, there are some particular pitfalls that should be avoided. The Raman spectra dataset should be large enough for the training and validating steps in order to make sure that the learning process is sufficient. In addition, collection of Raman spectral data is more important than models and algorithms themselves since over- or underrepresented data will lead to biased predictions. Moreover, what machine learning algorithms to choose and how the mode parameters are determined are also crucial for Raman spectral analysis.

RAMAN SPECTRAL DATABASE OF BACTERIAL PATHOGENS

A precondition for using machine learning to analyze Raman spectral data is a database with validated reference spectra of bacterial species and phenotypes (Moawad et al., 2019). It is convenient to measure single bacterial spectra from cultures in the lab, which is normally crucial to build a preliminary Raman spectral database. However, in order for the database to be functional in real-life environment, a database of Raman spectra from environmental or patient samples is required (Pahlow et al., 2015). Raman spectral databases have been constructed in a variety of fields, such as minerals, organics, inorganics, essential oils, pigments, and carbohydrates, which greatly facilitates the detection of these materials and further increases the applications of RS in corresponding fields (Strola et al., 2014; Kumar et al., 2015; El Mendili et al., 2019). Thus, a standard database of Raman spectra for bacterial pathogens would be very convenient and highly demanded in species identification and antibiotic resistance profiling.

Lorenz et al. (2017) emphasized the importance of Raman microscopic databases in the identification of leading pathogens in environmental and patient samples. Strola et al. (2014) constructed a reference database including a total of 1,200 spectra over seven bacterial species, based on which the success rate of bacterial species identification approaches 87% via SVM classification. Kloß et al. (2014) built up a Raman database containing 10,000 single-cell spectra for 34 bacterial strains belonging to 13 different species in ascitic fluids. In addition, Novelli-Rousseau et al. (2018) tried to distinguish antibiotic-resistant and antibiotic-susceptible *E. coli* based on a database with 3,668 Raman spectra. Some other studies also implement several small bacterial Raman databases that greatly promote the application of RS in bacterial analysis (Muhtar et al., 2016).

Unfortunately, at current stage, there is very little effort dedicating to the integration of small database into a large and standard Raman spectral database in bacterial field that may be widely used in different microbiological and clinical labs. A particular reason for such a deficiency is that Raman spectra from different studies are tailor-made and group specific, which greatly hinders data standardization (Lorenz et al., 2017). In order to facilitate the standardization of RS data, metadata annotation with minimal sample preparation and acquisition of Raman spectra is indispensable, which could not only alleviate technical noisy signals but also improve reproducibility in RS experiments (García-Timermans et al., 2018). Furthermore, sample preparation recommendations and data-processing guidelines should be introduced for future work, which shall greatly promote the application of RS and translate it into a routine diagnostic method in clinical laboratory.

SUMMARY

Raman spectroscopy can provide bacterial phenotypic information in details and vast amounts. Although numerous studies focus on rapid identification of bacterial species and antibiotic resistance profiles by RS, the real situation is that the technique has not been fully explored in clinical settings yet. Currently, most Raman spectra of bacterial pathogens are based on pure bacterial isolates, which heavily relies on medium culture, while Raman spectra from actual clinical samples are still rare. Recently, with the development of nanoparticles and nanostructured surfaces, SERS greatly improved the signal intensity of Raman spectra, which greatly contributes to a better differentiation of bacterial infections. In addition, Raman spectra consist of the spectra of a large set of complex chemical mixtures, requiring machine learning methods for data processing, such as artificial intelligence and deep learning, rather than classical linear methods. However, problems encountered during machine learning-assisted analysis involve overfitting or underfitting of the models due to the large dimension and small sample size problem of Raman spectra, although there are different dimension reduction methods like principal component analysis in use to address the issue. In addition, standard database of RS for bacterial pathogens is also a guarantee of the accurate and timely laboratory diagnosis when recruiting machine learning methods. In sum, the techniques of rapid and reliable automatic measurement of the Raman spectra of clinical samples from the real word are eagerly and urgently needed for the applicability of bacterial typing and antibiotic resistance profiling in clinical settings, which shall be achieved in foreseeable future with the fast development of novel Raman spectroscopic techniques, nanostructural materials, computational methods, and standardized databases.

AUTHOR CONTRIBUTIONS

LW conceptualized the manuscript. LW, XZ, and BG contributed to project administration. LW, WL, J-WT, Q-HL, BG, and XZ wrote the original manuscript. LW, WL, P-BW, and XZ provided platform, resources, and student supervision. J-JW, J-WT, M-MW, and Y-CP revised the manuscript. All authors read and approved the final manuscript.

FUNDING

LW would like to thank the financial support of National Natural Science Foundation of China (31900022), Natural Science Foundation of Jiangsu Province (BK20180997), Young Science and Technology Innovation Team of Xuzhou Medical University (TD202001) and Jiang-Su Qing-Lan Project (2020). BG thanks the financial support of National Natural Science Foundation of China (81871734, 81471994), Key R&D Program of Jiangsu Province (BE2020646), Jiangsu Provincial Medical Talent (ZDRCA2016053), Six Talent Peaks Project of Jiangsu Province (WSN-135), Advanced Health Talent of Six-one Project of Jiangsu Province (LGY2016042), and Research Foundation for Advanced Talents of Guangdong Provincial People's Hospital (KJ012021097).

ACKNOWLEDGMENTS

We are grateful for the reviewers in terms of their constructive comments and opinions, which greatly improves the quality and expand the views of this manuscript.

REFERENCES

Abayasekara, L. M., Perera, J., Chandrasekharan, V., Gnanam, V. S., Udunuwara, N. A., Liyanage, D. S., et al. (2017). Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. *BMC Infect. Dis.* 17:631. doi: 10.1186/s12879-017-2727-8

Ashton, L., Lau, K., Winder, C. L., and Goodacre, R. (2011). Raman spectroscopy: lighting up the future of microbial identification. *Future Microbiol.* 6, 991–997. doi: 10.2217/fmb.11.89

Balloux, F., Brønstad Brynildsrud, O., Van Dorp, L., Shaw, L. P., Chen, H., Harris, K. A., et al. (2018). From theory to practice: translating whole-genome sequencing (WGS) into the Clinic. *Trends Microbiol.* 26, 1035–1048. doi: 10.1016/j.tim.2018.08.004

Barghouthi, S. A. (2011). A universal method for the identification of bacteria based on general PCR primers. *Indian J. Microbiol.* 51, 430–444. doi: 10.1007/s12088-011-0122-5

Boardman, A. K., Wong, W. S., Premasiri, W. R., Ziegler, L. D., Lee, J. C., Miljkovic, M., et al. (2016). Rapid detection of bacteria from blood with surface-enhanced raman spectroscopy. *Anal. Chem.* 88, 8026–8035.

Bumbrah, G. S., and Sharma, R. M. (2016). Raman spectroscopy – basic principle, instrumentation and selected applications for the characterization of drugs of abuse. *Egypt. J. Forensic Sci.* 6, 209–215. doi: 10.1016/j.ejfs.2015.06.001

Burnham, C.-A. D., Leeds, J., Nordmann, P., O'grady, J., and Patel, J. (2017). Diagnosing antimicrobial resistance. *Nat. Rev. Microbiol.* 15, 697–703.

Carlomagno, C., Bertazioli, D., Gualerzi, A., Picciolini, S., Banfi, P. I., Lax, A., et al. (2021). COVID-19 salivary raman fingerprint: innovative approach for the detection of current and past SARS-CoV-2 infections. *Sci. Rep.* 11, 1–13.

Chang, K.-W., Cheng, H.-W., Shiue, J., Wang, J.-K., Wang, Y.-L., and Huang, N.-T. (2019). Antibiotic susceptibility test with surface-enhanced raman scattering in a microfluidic system. *Anal. Chem.* 91, 10988–10995. doi: 10.1021/acs.analchem.9b01027

Chen, K., Yuen, C., Aniweh, Y., Preiser, P., and Liu, Q. (2016). Towards ultrasensitive malaria diagnosis using surface enhanced Raman spectroscopy. *Sci. Rep.* 6, 1–10.

Chiu, C. Y., and Miller, S. A. (2019). Clinical metagenomics. *Nat. Rev. Genet.* 20, 341–355.

Cialla-May, D., Schmitt, M., and Popp, J. (2019). Theoretical principles of Raman spectroscopy. *Phys. Sci. Rev.* 4, 1–9. doi: 10.1515/9783110515312-001

de Siqueira E Oliveira, F. S., Da Silva, A. M., Pacheco, M. T. T., Giana, H. E., and Silveira, L. (2020). Biochemical characterization of pathogenic bacterial species using Raman spectroscopy and discrimination model based on selected spectral features. *Lasers Med. Sci.* 36, 289–302. doi: 10.1007/s10103-020-03028-9

Demirtaş, Ö, DoĐanay, D., Öztürk, I. M., Ünalan, H. E., and Bek, A. (2020). Facile preparation of nanoparticle based SERS substrates for trace molecule detection. *Phys. Chem. Chem. Phys.* 22, 21139–21146. doi: 10.1039/d0cp01866j

Deurenberg, R. H., Bathoorn, E., Chlebowicz, M. A., Couto, N., Ferdous, M., García-Cobos, S., et al. (2017). Application of next generation sequencing in clinical microbiology and infection prevention. *J. Biotechnol.* 243, 16–24.

Donald, C., Goh, B., Ching, K., Soares Magalhães, R. J., Ciocchetta, S., Edstein, M. D., et al. (2021). The application of spectroscopy techniques for diagnosis of malaria parasites and arboviruses and surveillance of mosquito vectors: a systematic review and critical appraisal of evidence. *PLoS Negl. Trop. Dis.* 15:e0009218. doi: 10.1371/journal.pntd.0009218

Eberhardt, K., Stiebing, C., Matthäus, C., Schmitt, M., and Popp, J. (2015). Advantages and limitations of Raman spectroscopy for molecular diagnostics: an update. *Exp. Rev. Mol. Diagn.* 15, 773–787. doi: 10.1586/14737159.2015.1036744

El Mendili, Y., Vaitkus, A., Merkys, A., Gražulis, S., Chateigner, D., Mathevet, F., et al. (2019). Raman open database: first interconnected Raman–X-ray diffraction open-access resource for material identification. *J. Appl. Crystallogr.* 52, 618–625. doi: 10.1107/s1600576719004229

Florio, W., Tavanti, A., Barnini, S., Ghelardi, E., and Lupetti, A. (2018). Recent advances and ongoing challenges in the diagnosis of microbial infections by MALDI-TOF mass spectrometry. *Front. Microbiol.* 9:1097. doi: 10.3389/fmicb.2018.01097

Fournier, P.-E., Dubourg, G., and Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med.* 6, 1–15.

Franco-Duarte, R., Èernáková, L., Kadam, S. S., Kaushik, K., Salehi, B., Bevilacqua, A., et al. (2019). Advances in chemical and biological methods to identify microorganisms—from past to present. *Microorganisms* 7, 1–32.

Galvan, D. D., and Yu, Q. (2018). Surface-enhanced raman scattering for rapid detection and characterization of antibiotic-resistant bacteria. *Adv. Healthcare Mater.* 7, 1–27.

García-Timermans, C., Rubbens, P., Kerckhof, F.-M., Buysschaert, B., Khalenkow, D., Waegeman, W., et al. (2018). Label-free Raman characterization of bacteria calls for standardized procedures. *J. Microbiol. Methods* 151, 69–75. doi: 10.1016/j.mimet.2018.05.027

Han, Y.-Y., Lin, Y.-C., Cheng, W.-C., Lin, Y.-T., Teng, L.-J., Wang, J.-K., et al. (2020). Rapid antibiotic susceptibility testing of bacteria from patients' blood via assaying bacterial metabolic response with surface-enhanced Raman spectroscopy. *Sci. Rep.* 10, 1–18.

Hou, T.-Y., Chiang-Ni, C., and Teng, S.-H. (2019). Current status of MALDI-TOF mass spectrometry in clinical microbiology. *J. Food Drug Anal.* 27, 404–414.

Jones, R. R., Hooper, D. C., Zhang, L., Wolverson, D., and Valev, V. K. (2019). Raman techniques: fundamentals and frontiers. *Nanoscale Res. Lett.* 14, 1–34. doi: 10.1063/9780735422209_007

Khan, Z. A., Siddiqui, M. F., and Park, S. (2019). Current and emerging methods of antibiotic susceptibility testing. *Diagnostics* 9, 1–17.

Kloß, S., Rösch, P., Pfister, W., Kiehntopf, M., and Popp, J. (2014). Toward culture-free raman spectroscopic identification of pathogens in ascitic fluid. *Anal. Chem.* 87, 937–943. doi: 10.1021/ac503373r

Krafft, C., and Popp, J. (2014). Raman-based technologies for biomedical diagnostics. *Compr. Biomed. Phys.* 4, 189–208. doi: 10.1016/b978-0-444-53632-7.00415-9

Kubina, R., and Dziedzic, A. (2020). Molecular and serological tests for COVID-19. A comparative review of SARS-CoV-2 coronavirus laboratory and point-of-care diagnostics. *Diagnostics* 10, 1–18.

Kumar, V., Kampe, B., Rösch, P., and Popp, J. (2015). Classification and identification of pigmented cocci bacteria relevant to the soil environment via Raman spectroscopy. *Environ. Sci. Pollut. Res.* 22, 19317–19325. doi: 10.1007/s11356-015-4593-5

Kwong, J. C., Mccallum, N., Sintchenko, V., and Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology* 47, 199–210. doi: 10.1097/pat.0000000000000235

Lai, C.-C., Wang, J.-H., and Hsueh, P.-R. (2020). Population-based seroprevalence surveys of anti-SARS-CoV-2 antibody: an up-to-date review. *Int. J. Infectious Dis.* 101, 314–322. doi: 10.1016/j.ijid.2020.10.011

Lane, L. A., Qian, X., and Nie, S. (2015). SERS nanoparticles in medicine: from label-free detection to spectroscopic tagging. *Chem. Rev.* 115, 10489–10529. doi: 10.1021/acs.chemrev.5b00265

Law, J. W.-F., Ab Mutalib, N.-S., Chan, K.-G., and Lee, L.-H. (2015). Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Front. Microbiol.* 5:770. doi: 10.3389/fmicb.2014.00770

Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., et al. (2020). Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *mBio* 11, 1–22.

Li, J., Wang, C., Shi, L., Shao, L., Fu, P., Wang, K., et al. (2019). Rapid identification and antibiotic susceptibility test of pathogens in blood based on magnetic separation and surface-enhanced Raman scattering. *Microchim. Acta* 186, 1–12.

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21, 597–614. doi: 10.1038/s41576-020-0236-x

Lorenz, B., Wichmann, C., Stöckel, S., Rösch, P., and Popp, J. (2017). Cultivation-free raman spectroscopic investigations of bacteria. *Trends Microbiol.* 25, 413–424. doi: 10.1016/j.tim.2017.01.002

Lu, X., Huang, Q., Miller, W. G., Aston, D. E., Xu, J., Xue, F., et al. (2012). Comprehensive detection and discrimination of campylobacter species by use of confocal micro-raman spectroscopy and multilocus sequence typing. *J. Clin. Microbiol.* 50, 2932–2946. doi: 10.1128/jcm.01144-12

Lussier, F., Thibault, V., Charron, B., Wallace, G. Q., and Masson, J.-F. (2020). Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends Anal. Chem.* 124, 1–15.

Maquelin, K., Kirschner, C., Choo-Smith, L. P., Ngo-Thi, N. A., Van Vreeswijk, T., StäMmler, M., et al. (2003). Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures. *J. Clin. Microbiol.* 41, 324–329. doi: 10.1128/jcm.41.1.324-329.2003

Marques, A., Veigas, B., Araújo, A., Pagará, B., Baptista, P. V., Águas, H., et al. (2019). Paper-based SERS platform for one-step screening of tetracycline in milk. *Sci. Rep.* 9, 1–8.

Moawad, A. A., Silge, A., Bocklitz, T., Fischer, K., Rösch, P., Roesler, U., et al. (2019). A machine learning-based raman spectroscopic assay for the identification of *Burkholderia mallei* and related species. *Molecules* 24, 1–15.

Muhtar, I., Mengyue, G., Fang, P., Aiguo, S., and Jiming, H. (2016). Discrimination of natural gas-related bacteria by means of micro-Raman spectroscopy. *Vib. Spectrosc.* 82, 44–49. doi: 10.1016/j.vibspec.2015.11.006

Novelli-Rousseau, A., Espagnon, I., Filiputti, D., Gal, O., Douet, A., Mallard, F., et al. (2018). Culture-free antibiotic-susceptibility determination from single-bacterium Raman Spectra. *Sci. Rep.* 8, 1–12.

Ouyang, L., Ren, W., Zhu, L., and Irudayaraj, J. (2017). Prosperity to challenges: recent approaches in SERS substrate fabrication. *Rev. Anal. Chem.* 36, 1–22.

Pahlow, S., Meisel, S., Cialla-May, D., Weber, K., Rösch, P., and Popp, J. (2015). Isolation and identification of bacteria by means of Raman spectroscopy. *Adv. Drug Delivery Rev.* 89, 105–120. doi: 10.1016/j.addr.2015.04.006

Rebrošová, K., Šiler, M., Samek, O., Rùžièka, F., Bernatová, S., Holá, V., et al. (2017). Rapid identification of staphylococci by Raman spectroscopy. *Sci. Rep.* 7, 1–8.

Rossen, J. W. A., Friedrich, A. W., and Moran-Gilad, J. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin. Microbiol. Infection* 24, 355–360. doi: 10.1016/j.cmi.2017.11.001

Sakamoto, S., Putalun, W., Vimolmangkang, S., Phoolcharoen, W., Shoyama, Y., Tanaka, H., et al. (2017). Enzyme-linked immunosorbent assay for the quantitative/qualitative analysis of plant secondary metabolites. *J. Nat. Med.* 72, 32–42. doi: 10.1007/s11418-017-1144-z

Sauer, S., and Kliem, M. (2010). Mass spectrometry tools for the classification and identification of bacteria. *Nat. Rev. Microbiol.* 8, 74–82. doi: 10.1038/nrmicro2243

Sil, S., Mukherjee, R., Kumar, N. S., Umapathy, S., Popp, J., and Gergely, C. (2020). "Potential and challenges of pathogen detection using Raman spectroscopy," in *Proceedings of the Biomedical Spectroscopy, Microscopy, and Imaging* (SPIE Photonics Europe).

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* 6:791. doi: 10.3389/fmicb.2015.00791

Smith, R., Wright, K. L., and Ashton, L. (2016). Raman spectroscopy: an evolving technique for live cell studies. *Analyst* 141, 3590–3600. doi: 10.1039/c6an00152a

Solís, D. M., Taboada, J. M., Obelleiro, F., Liz-Marzán, L. M., and García De Abajo, F. J. (2017). Optimization of nanoparticle-based sers substrates through large-scale realistic simulations. *ACS Photonics* 4, 329–337. doi: 10.1021/acsphotonics.6b00786

Stratakos, A., Linton, M., Ward, P., Ijaz, U., Scates, P., Mcbride, J., et al. (2019). Integrated phenotypic and genomics analysis to elucidate differences in stress resistance and virulence of *Listeria monocytogenes* strains. *Access Microbiol.* 1, 1–1.

Strola, S. A., Baritaux, J.-C., Schultz, E., Simon, A. C., Allier, C., Espagnon, I., et al. (2014). Single bacteria identification by Raman spectroscopy. *J. Biomed. Optics* 19, 1–13.

Tien, N., Lin, T.-H., Hung, Z.-C., Lin, H.-S., Wang, I. K., Chen, H.-C., et al. (2018). Diagnosis of bacterial pathogens in the urine of urinary-tract-infection patients using surface-enhanced raman spectroscopy. *Molecules* 23, 1–14.

Uysal Ciloglu, F., Saridag, A. M., Kilic, I. H., Tokmakci, M., Kahraman, M., and Aydin, O. (2020). Identification of methicillin-resistant Staphylococcus aureus bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. *Analyst* 145, 7559–7570. doi: 10.1039/d0an00476f

van Sorge, N. M., and Doran, K. S. (2012). Defense at the border: the blood–brain barrier versus bacterial foreigners. *Future Microbiol.* 7, 383–394. doi: 10.2217/fmb.12.1

Váradi, L., Luo, J. L., Hibbs, D. E., Perry, J. D., Anderson, R. J., Orenga, S., et al. (2017). Methods for the detection and identification of pathogenic bacteria: past, present, and future. *Chem. Soc. Rev.* 46, 4818–4832. doi: 10.1039/c6cs00693k

Verroken, A., Defourny, L., Lechgar, L., Magnette, A., Delmée, M., and Glupczynski, Y. (2014). Reducing time to identification of positive blood cultures with MALDI-TOF MS analysis after a 5-h subculture. *Eur. J. Clin. Microbiol. Infect. Dis.* 34, 405–413. doi: 10.1007/s10096-014-2242-4

Wang, K., Li, S., Petersen, M., Wang, S., and Lu, X. (2018). Detection and characterization of antibiotic-resistant bacteria using surface-enhanced raman spectroscopy. *Nanomaterials* 8, 1–21.

Wang, W., Mcgregor, H., Short, M., and Zeng, H. (2016). Clinical utility of Raman spectroscopy: current applications and ongoing developments. *Adv. Health Care Technol.* 2, 13–29. doi: 10.2147/ahct.s96486

Wei, D., Chen, S., and Liu, Q. (2015). Review of fluorescence suppression techniques in Raman Spectroscopy. *Appl. Spectrosc. Rev.* 50, 387–406. doi: 10.1080/05704928.2014.999936

Xu, N., Wang, W., Chen, F., Li, W., and Wang, G. (2020). ELISA is superior to bacterial culture and agglutination test in the diagnosis of brucellosis in an endemic area in China. *BMC Infect. Dis.* 20:11. doi: 10.1186/s12879-019-4729-1

Yeh, Y.-T., Gulino, K., Zhang, Y., Sabestien, A., Chou, T.-W., Zhou, B., et al. (2020). A rapid and label-free platform for virus capture and identification from clinical samples. *Proc. Natl. Acad. Sci.U.S.A.* 117, 895–901. doi: 10.1073/pnas.1910113117

Yin, G., Li, L., Lu, S., Yin, Y., Su, Y., Zeng, Y., et al. (2021). An efficient primary screening of COVID−19 by serum Raman spectroscopy. *J. Raman Spectrosc.* 52, 949–958. doi: 10.1002/jrs.6080

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.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fmicb.2021.649969

BacAnt: A Combination Annotation Server for Bacterial DNA Sequences to Identify Antibiotic Resistance Genes, Integrons, and Transposable Elements

Xiaoting Hua[1,2,3†], Qian Liang[2,4†‡], Min Deng[5†], Jintao He[1,2,3], Meixia Wang[2,4], Wenjie Hong[2,4], Jun Wu[6], Bian Lu[7], Sebastian Leptihn[1,8], Yunsong Yu[1,2,3] and Huan Chen[1,2,4*‡]*

[1] Department of Infectious Diseases, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2] Key Laboratory of Microbial Technology and Bioinformatics of Zhejiang Province, Zhejiang Institute of Microbiology, Hangzhou, China, [3] Regional Medical Center for National Institute of Respiratory Diseases, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China, [4] National Medical Products Administration Key Laboratory for Testing and Risk Warning of Pharmaceutical Microbiology, Zhejiang Institute of Microbiology, Hangzhou, China, [5] Department of Infectious Diseases, The First Hospital of Jiaxing, The First Affiliated Hospital of Jiaxing University, Jiaxing, China, [6] Lin'an Center for Disease Control and Prevention, Lin'an, China, [7] Xiaoshan Center for Disease Control and Prevention, Hangzhou, China, [8] Zhejiang University-University of Edinburgh Institute, School of Medicine, Zhejiang University, Hangzhou, China

Whole genome sequencing (WGS) of bacteria has become a routine method in diagnostic laboratories. One of the clinically most useful advantages of WGS is the ability to predict antimicrobial resistance genes (ARGs) and mobile genetic elements (MGEs) in bacterial sequences. This allows comprehensive investigations of such genetic features but can also be used for epidemiological studies. A plethora of software programs have been developed for the detailed annotation of bacterial DNA sequences, such as rapid annotation using subsystem technology (RAST), Resfinder, ISfinder, INTEGRALL and The Transposon Registry. Unfortunately, to this day, a reliable annotation tool of the combination of ARGs and MGEs is not available, and the generation of genbank files requires much manual input. Here, we present a new webserver which allows the annotation of ARGs, integrons and transposable elements at the same time. The pipeline generates genbank files automatically, which are compatible with Easyfig for comparative genomic analysis. Our BacAnt code and standalone software package are available at https://github.com/xthua/bacant with an accompanying web application at http://bacant.net.

Keywords: BacAnt, annotation, antibiotic resistance gene, integron, transposable element

OPEN ACCESS

Edited by:
*Liang (Leon) Wang,
Xuzhou Medical University, China*

Reviewed by:
*Fani Chatzopoulou,
Aristotle University of Thessaloniki,
Greece
Trinad Chakraborty,
University of Giessen, Germany*

***Correspondence:**
*Yunsong Yu
yvys119@zju.edu.cn
Huan Chen
10607068@zju.edu.cn*

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

[‡]Present address:
*Qian Liang and Huan Chen,
Hangzhou Digital Micro Biotech Co.,
Ltd., Hangzhou, China*

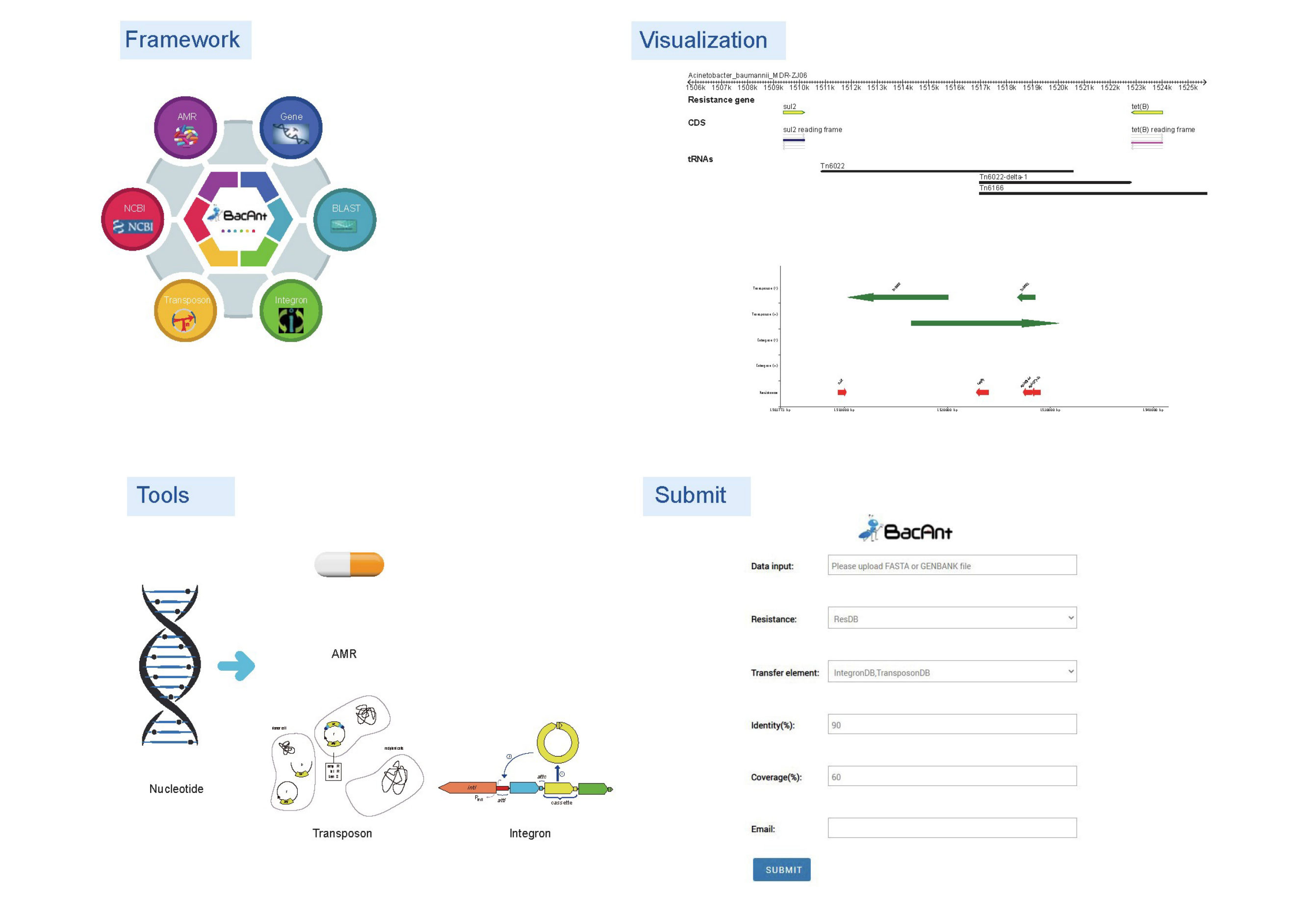
Specialty section:
*This article was submitted to
Systems Microbiology,
a section of the journal
Frontiers in Microbiology*

Received: *06 January 2021*
Accepted: *31 May 2021*
Published: *23 July 2021*

Citation:
Hua X, Liang Q, Deng M, He J, Wang M, Hong W, Wu J, Lu B, Leptihn S, Yu Y and Chen H (2021) BacAnt: A Combination Annotation Server for Bacterial DNA Sequences to Identify Antibiotic Resistance Genes, Integrons, and Transposable Elements. Front. Microbiol. 12:649969. doi: 10.3389/fmicb.2021.649969

INTRODUCTION

The era of next-generation sequencing (NGS) took off in 2005 with the commercial release of massively parallel pyrosequencing (Margulies et al., 2005). The NGS technology developed rapidly in the past years and has made substantial improvements in terms of quality and yield. With the rapid decrease of sequencing costs, falling by as much as 80% year over year, whole genome



Data input: Please upload FASTA or GENBANK file

Resistance: ResDB

Transfer element: IntegronDB,TransposonDB

Identity(%): 90

Coverage(%): 60

Email:

GRAPHICAL ABSTRACT | Overall BacAnt's service workflow. The figure is adapted with permissions from IntegronFinder (Cury et al., 2016).

sequencing (WGS) of bacteria has become a routine method in diagnostic laboratories (Didelot et al., 2012). NGS applications include WGS, targeted NGS and metagenomic NGS. Among them, the most common use of WGS is for simultaneous identification, typing, and/or antimicrobial susceptibility prediction of pathogens (Mitchell and Simner, 2019). One of the most exciting advantages of NGS is the ability to predict antimicrobial resistance genes (ARGs) and mobile genetic elements (MGEs) in bacteria, which allows the investigation of both, the organization and structure of such genetic features, and the epidemiology for the distribution of bacterial strains or virulence genes, including the spread and distribution of antibiotic-resistant bacteria as part of surveillance programs (Zhou et al., 2015; Mitchell and Simner, 2019). Every day, a massive number of bacterial genomes is being sequenced using NGS technology in laboratories across the globe, with genomes released at remarkable rates. With this huge amount of data available, it is important to extract project-relevant information easily. However, in publicly available databases, most of bacterial genomes are available as contigs which have been constructed employing auto-annotation algorithms. Over the years, highly efficient methods for bacterial genome annotation have been developed that do not require much user input.

Rapid Annotation using Subsystem Technology (RAST) is a widely used webserver for genome annotations of microbial species (Aziz et al., 2008). Although the performance using RAST-based annotation is very useful, several important limitations remain. For example, RAST will label many Open Reading Frames (ORFs) as "hypothetical proteins," and the performance to identify ARGs and label them as such, is fairly limited as the algorithm is not tailored toward this purpose. Based on the RAST system, the Pathosystems Resource Integration Center (PATRIC) improved the data collection of ARGs, and provided users a more powerful analysis for both genomes and individual genes (Wattam et al., 2018). Another available annotation server is Resfinder which is managed by the Center for Genomic Epidemiology; it provides a convenient way of identifying acquired ARGs in sequenced bacterial isolates (Zankari et al., 2012). In addition to annotations of ARGs, some databases specifically designed to annotate MGEs such as insertion sequences (ISfinder), integrons (INTEGRALL) and transposable elements (The Transposon Registry) have been created (Siguier et al., 2006; Moura et al., 2009; Tansirichaiya et al., 2019). ISs are abundant mobile elements in bacteria, which are responsible for the mobilization of many genes, including those mediating ARG (Razavi et al., 2020). Such

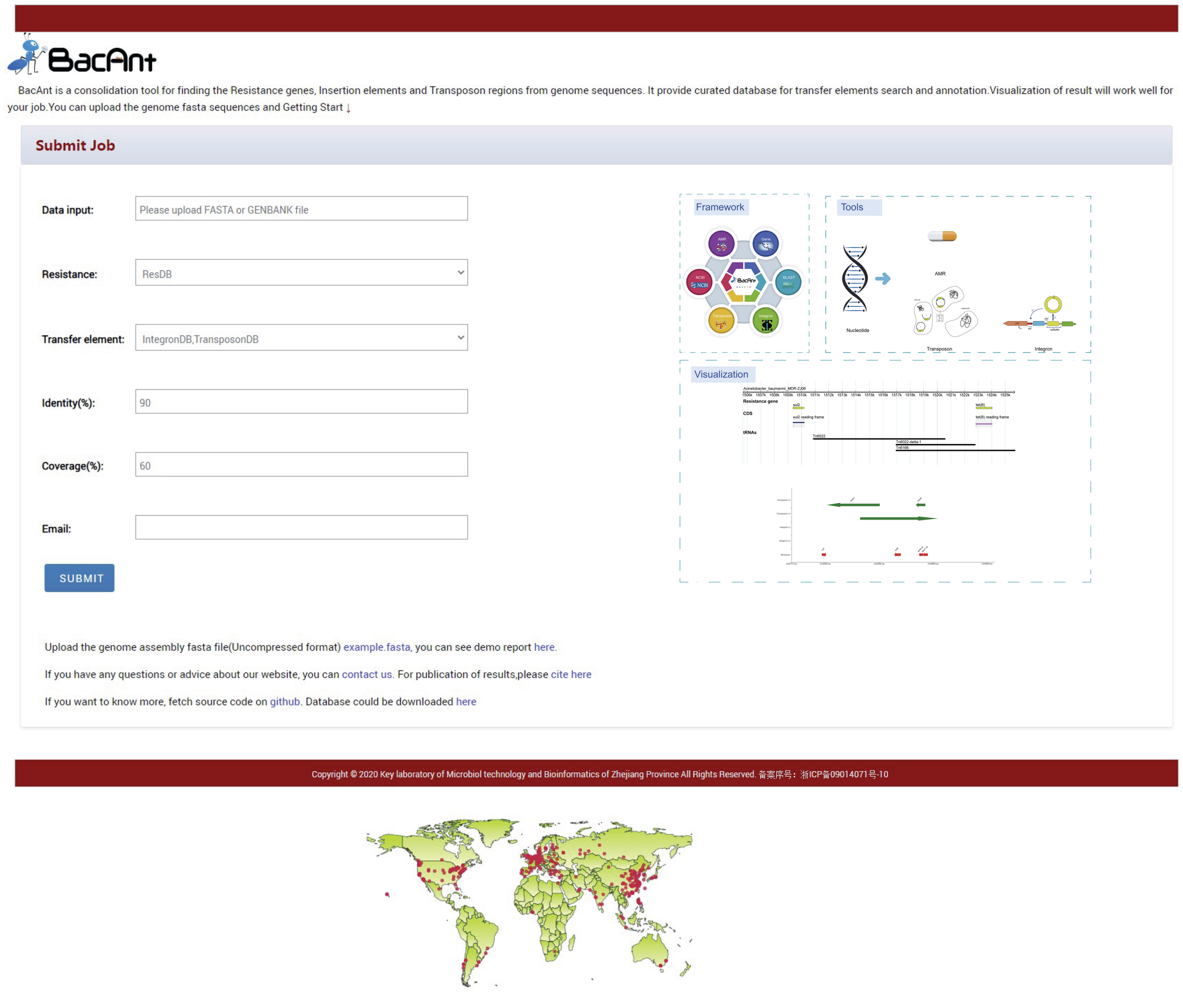
ARGs are often found in the genetic context of specific ISs, while ISs flanking regions are diverse (Razavi et al., 2020). For example, a clear association of ARGs with class 1 integrons can be observed (Partridge et al., 2018). The analysis of which ISs are associated with ARG genes would help to discover novel AMGs (Razavi et al., 2020). In addition, there is a major interest to explore how ARGs spread via MGEs (Che et al., 2021). The early identification of ARGs in bacteria would facilitate surveillance and molecular diagnostics (Razavi et al., 2020). Also, inter/intra-species genetic transfer events of MGEs are responsible for the emergence and rapid spread of resistance (Subedi et al., 2018). Therefore, the knowledge of MGE-associated drug resistance is crucial for the monitoring of resistance in microbial species. Unfortunately, up to now, a rapid annotation tool of the combination of ARGs and MGEs is not available, and the generation of genbank files has to be done manually. Therefore, we created a new program/pipeline called BacAnt, which rapidly and efficiently annotates ARGs, integrons, and transposable elements in a single step and generates a genbank file automatically which is compatible with the Easyfig program for comparative genomic analysis.

MATERIALS AND METHODS

Reference Sequences

Three curated databases for BacAnt tool are used, including ResDB (resistance gene sequence database), IntegronDB (integron sequence database) and TransposonDB (transposon sequence database). We collected 5029 sequences from NCBI Bacterial Antimicrobial Resistance Reference Gene Database (PRJNA313047[1], version: 2019-09-06.1) into ResDB at 2019-12-01. In addition, we collected 1094 sequences from

[1] https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047



BacAnt is a consolidation tool for finding the Resistance genes, Insertion elements and Transposon regions from genome sequences. It provide curated database for transfer elements search and annotation.Visualization of result will work well for your job.You can upload the genome fasta sequences and Getting Start ↓

Submit Job

Data input: Please upload FASTA or GENBANK file

Resistance: ResDB

Transfer element: IntegronDB,TransposonDB

Identity(%): 90

Coverage(%): 60

Email:

Upload the genome assembly fasta file(Uncompressed format) example.fasta, you can see demo report here.

If you have any questions or advice about our website, you can contact us. For publication of results,please cite here

If you want to know more, fetch source code on github. Database could be downloaded here

Copyright © 2020 Key laboratory of Microbiol technology and Bioinformatics of Zhejiang Province All Rights Reserved. 备案序号：浙ICP备09014071号-10

FIGURE 1 | Screenshots of the BacAnt web interface. Users upload an assembled file from their local personal computer and select the desired annotation database. "Framework" lists the public database and tools integrated in BacAnt; "Tools" functions for whole genome sequence annotation based on user uploaded sequence(s); "Visualization" allows to visualize the annotation results for uploaded sequence(s).

INTEGRALL (version: 2017-11-30)[2] to be included into IntegronDB by 2019-12-01. We also collected 234 sequences from THE TRANSPOSON REGISTRY (version: 2019-07-23)[3] into the TransposonDB by 2019-12-01.

Program for Identifying Antibiotic Resistance Genes and Mobile Elements

We created a python program (BacAnt) to identify ARGs and MEGs for bacteria nucleotide sequences with BLAST analysis. We first used the BLASTN program comparing input sequences with the reference database with an *e*-value 10^{-5}. For the detection of integrons, we used Integron_Finder to predict possible integrons and used the BLASTN program comparing the integron sequence with integronDB database for the best match sequence (Cury et al., 2016). We then filtered the raw results by identities and coverage (blast align match length/subject length). All results that pass the identity and coverage filter are retained for further analysis. The default threshold was set to 90% for identities and to 60% for coverage. Finally, we display the filtered results in text and genbank format, while also providing a visual output. Three types of annotations for the sequences are displayed in the same figure to guide the analysis of the genomic sequence.

BacAnt has six parameters. The user has two choices regarding the input sequence file: –nucleotide (–n), fasta format or –genbank (–g), genbank format. The required output path: –resultdir (–o). –databases (–d), reference databases, select all by default. –coverages (–c), coverage threshold, 60% by default. –identities (–i), identities threshold, 90% by default. In average, it takes about 2 min for each run (number of available cores: 6; threads: 24; memory 64G).

Website for BacAnt

For the analysis to be performed online, we developed a website which we call http://bacant.net. The pipeline running on the server is based on Python/Django, which allows the user to upload sequence files for the rapid identification of ARGs and MGEs. The output format allows the display of graphic

[2] http://integrall.bio.ua.pt

[3] https://transposon.lstmed.ac.uk

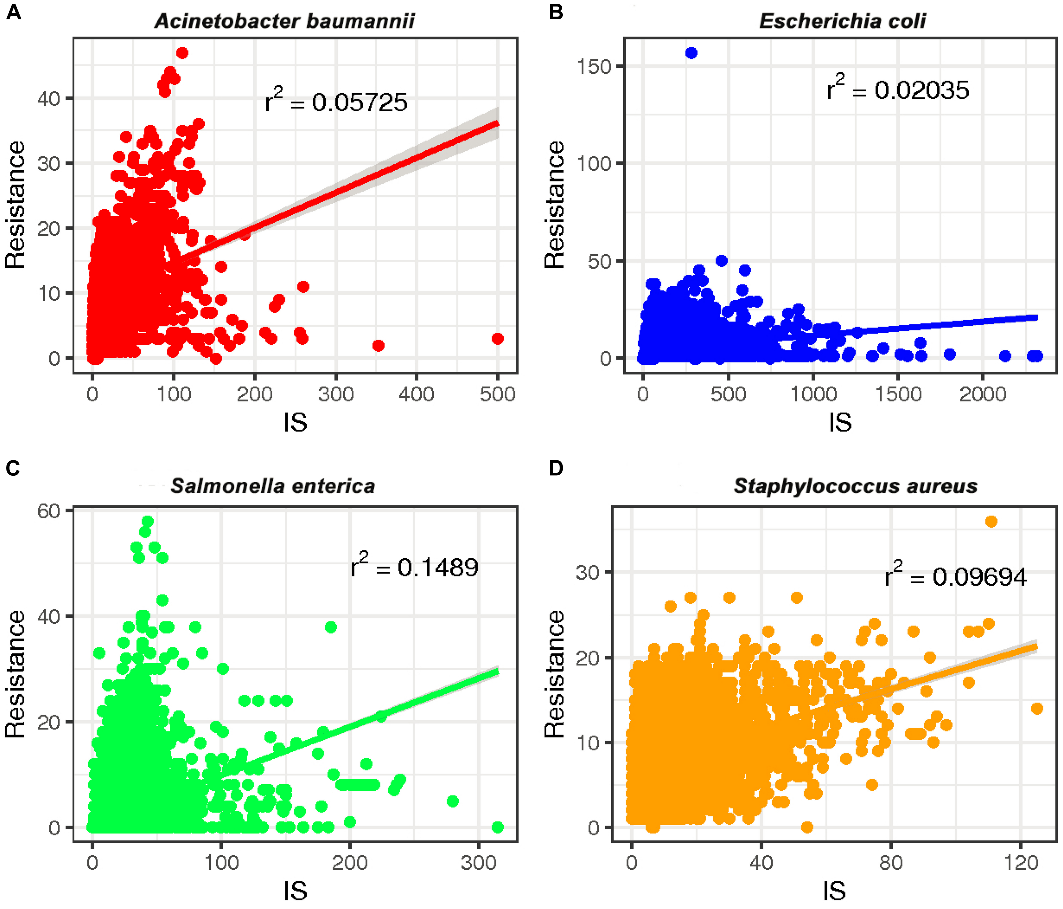


FIGURE 2 | The relationship between the number of drug-resistance genes (ARGs) and insertion sequences (IS) from four species: **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica*, and **(D)** *S. aureus*. The number of ARGs and ISs of each sample were extracted from the results of the BacAnt analysis, and the number was plotted as the horizontal and vertical coordinates, respectively. The scatter plot was created using the ggplot2 package in R (V3.6.2), and a trend line generated.

representations of the results. A demo report can be seen here: http://bacant.net/BacAnt/demo.

Datasets for Validation of BacAnt

BacAnt was validated with 1100 genomes (**Supplementary Table 1**) from eight species (*Acinetobacter baumannii*, *Bacillus cereus*, *Clostridioides difficile*, *Escherichia coli*, *Listeria monocytogenes*, *Salmonella enterica*, *Staphylococcus aureus*, and *Vibrio parahaemolyticus*). The BacAnt output was analyzed and compared with the results of NCBI AMRFinder. The parameters of BacAnt used in the study were: identity 0.9, coverage 0.6.

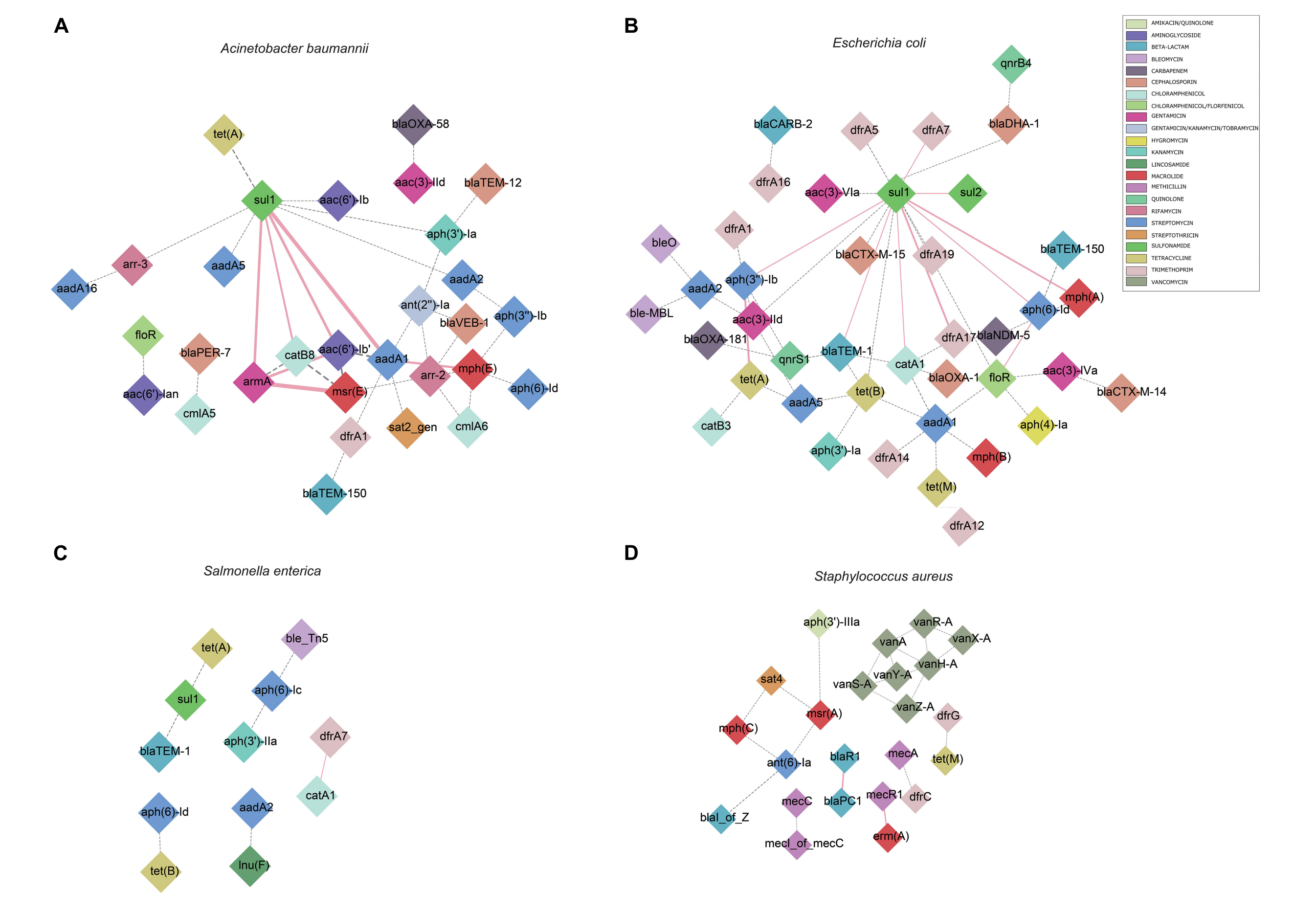
Examples Analysis Using BacAnt

The genome sequences of four species downloaded from NCBI, including *A. baumannii* (2019.11.17), *E. coli* (2019.5.7), *S. enterica* (2019.6.4), and *S. aureus* (2019.6.4) were used to illustrate the capabilities of our program. The accession number of the genomes used in this study were listed in **Supplementary Tables 2,3**. The raw sequence data were downloaded from the European Nucleotide Archive[4]. Sequence quality was assessed via FastQC v0.11.5[5], and low-quality sequence data and the adapter sequences were removed with Trimmomatic v0.36 (Bolger et al., 2014). The SPAdes software tool v3.11.0 was used to generate assembled genome with default parameter (Bankevich et al., 2012). The number of ARGs and MGEs of each sample were identified in the BacAnt analysis and plotted as horizontal and vertical coordinates, respectively. Scatter plots were created using the ggplot2 package in R (V3.6.2), together with the trend line (Wickham, 2016).

We created a network diagram by Cytoscape (v3.7.2) using the analysis results of BacAnt from four species genome sequences from NCBI (Shannon et al., 2003). Only ARG pairs with a distance of less than 10 kb and frequency of occurrence no less than 10 are extracted from the data to construct the network map. In addition, we also created a map from ARGs and insertion

[4] http://www.ebi.ac.uk/ena

[5] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



▭	AMIKACIN/QUINOLONE
▭	AMINOGLYCOSIDE
▭	BETA-LACTAM
▭	BLEOMYCIN
▭	CARBAPENEM
▭	CEPHALOSPORIN
▭	CHLORAMPHENICOL
▭	CHLORAMPHENICOL/FLORFENICOL
▭	GENTAMICIN
▭	GENTAMICIN/KANAMYCIN/TOBRAMYCIN
▭	HYGROMYCIN
▭	KANAMYCIN
▭	LINCOSAMIDE
▭	MACROLIDE
▭	METHICILLIN
▭	QUINOLONE
▭	RIFAMYCIN
▭	STREPTOMYCIN
▭	STREPTOTHRICIN
▭	SULFONAMIDE
▭	TETRACYCLINE
▭	TRIMETHOPRIM
▭	VANCOMYCIN

FIGURE 3 | Genomic proximity of antibiotic resistance genes (ARGs) in four species: **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica*, **(D)** *S. aureus*. ARG pairs with a distance of less than 10 kb and frequency of occurrence no less than 10 were screened to construct the network. Diamonds are used to represent ARGs, solid lines are used to represent gene pairs whose frequency is not less than 80% of the total number and labeled as pink, dotted lines represent gene pairs with frequency less than 80% of the total number and are marked as gray, and the width of the line indicates the frequency. Different antibiotics are displayed in different colors, corresponding to the ARGs.

sequence pairs with a distance of less than 10 kb and the frequency of occurrence no less than 10.

RESULTS AND DISCUSSION

BacAnt is a browser-based platform to annotate DNA sequences, and to visualize the annotation results. When using the web interface, the user first has the choice to upload DNA sequences as Fasta, Seq or GenBank files (**Figure 1**). The user then has the option to select one or multiple databases, which include ResDB, IntegronDB or TransponsonDB for sequence annotation. After the DNA sequence is submitted, the python-based program BacAnt will identify ARGs and MGEs in the bacterial nucleotide sequence.

The output of BacAnt commences with a summary of the annotation, followed by up to three tables including the annotation result from each database; should they have been selected in the first step. The final part of the BacAnt output visualizes an annotation result which is combined from all three databases. All annotation results obtained by running BacAnt, including figures and genbank files, can then be downloaded. The genebank files generated by BacAnt are compatible with Easyfig (Sullivan et al., 2011). As an example we used *A. baumannii* MDR-ZJ06 (NC_017171.2) to display the result of the annotation by BacAnt (**Figure 1**). The annotation output of the MDR-ZJ06 strain shows that the isolate harbors 19 ARGs, 124 integrons and 17 transposons.

BacAnt was validated with NCBI AMRFinder using 1,100 selected genomes. The file output "AMR.possible.tsv" in the BacAnt result was used for analysis in NCBI AMRFinder to test which of the programs is able to identify a larger number of ARGs with high accuracy. Both programs reported similar results regarding the number of resistance genes (**Supplementary Table 4**). However, the number of ARGS in BacAnt is slightly larger than that of NCBI AMRFinder. Some resistance genes were absent in the output of NCBI AMRFinder: aac(6′)-Iaa (NC_003197, aminoglycoside N-acetyltransferase) in *S. enterica*, bla_{EC-15} (NG_049081,class C extended-spectrum beta-lactamase EC-15) in *E. coli*, BcII (NG_056058, BcII family subclass B1 metallo-beta-lactamase in *B. cereus*.

To investigate whether a relationship between ARGs and ISs exists, we extracted the number of ARGs and ISs from the results of the BacAnt analysis of four species genome sequences from NCBI. The results show that the number of ARGs does not correlate with the number of ISs in the four species (Pearson's $R^2 < 0.8$, **Figure 2**). Previously, it was reported that at least eight MGEs were detected together with ARGs in *A. baumannii* (Leal et al., 2020). The plasmids were grouped into three categories based on the DNA transfer machinery:

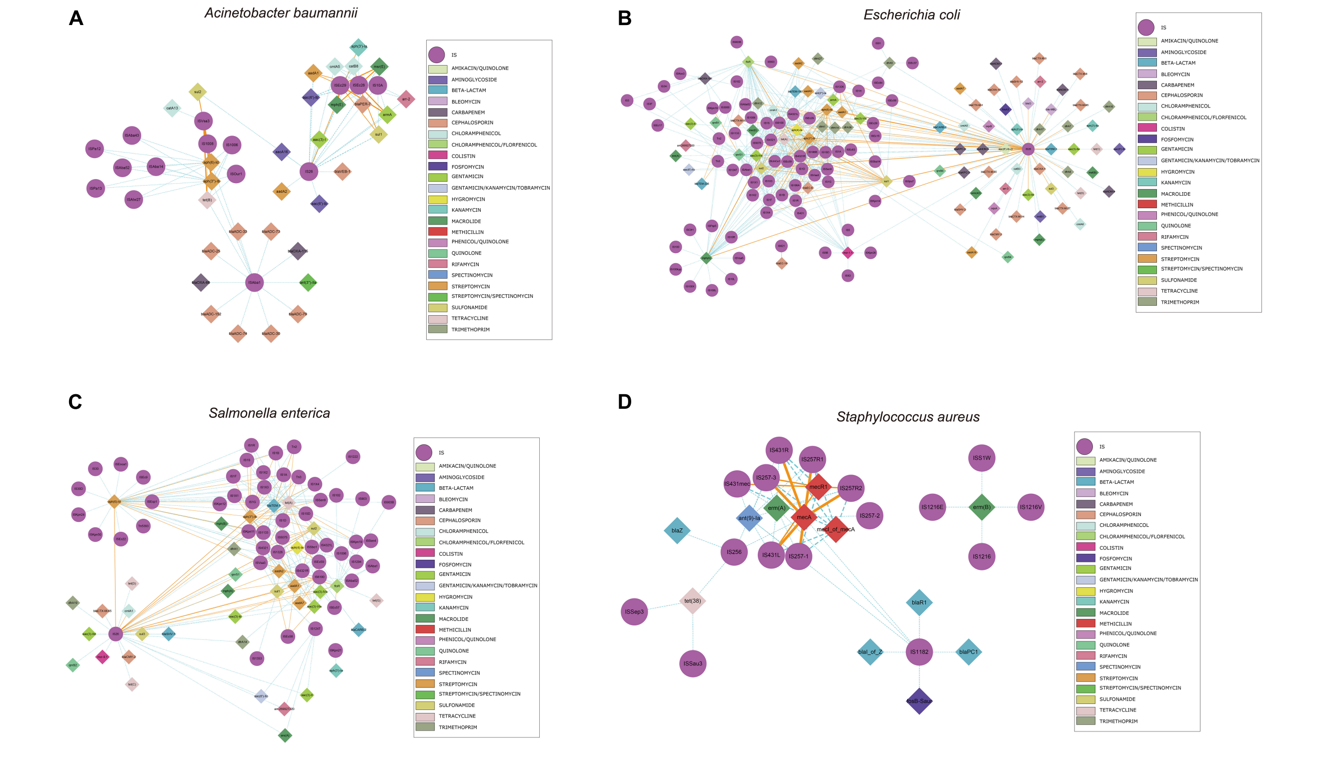


FIGURE 4 | Genomic proximity of antibiotic resistance genes (ARGs) and Insertion sequences (ISs) in four species: **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica,* and **(D)** *S. aureus*. ARGs and ISs pairs with a distance of less than 10 kb and frequency of occurrence no less than 10 were screened to construct the network. Diamond symbols are used to represent ARGs, circles for ISs. The solid line is the gene pair whose frequency is not less than 80% of the total number and labeled as orange, the dotted line represents the gene pair whose frequency is less than 80% of the total number and labeled as sky blue, and the width of the line indicates the frequency. Different colors are assigned to the antibiotics corresponding to the drug-resistant genes, the insertion sequences are displayed in gray.

conjugative, mobilizable and non-mobilizable (Smillie et al., 2010). Che et al. (2021) showed that most ARG genes are located in conjugative plasmids, which -together with ISs- play the most important role in mediating the horizontal transfer of ARGs. When the relationship between ARG and ISs were investigated, we did not analyze the location of ARGs which might explain why no significant correlation between the two was observed in our study.

We also calculated the pairwise association between ISs and ARGs. The results were subjected to a permutation test to differentiate between statistically significant associations and random chance (Razavi et al., 2020). Only statistically significant associations ($P < 0.001$) of ISs and ARGs were analyzed (**Supplementary Table 5**). We identified commonly found ARG pairs that were in close proximity to ISs (<10 kb apart), which allows the detection of gene cassettes that may play an important role in evolution, regulation and ARG exchange. ARG cassettes are generally small (2–7) and specific to the species we investigated (**Figure 3**). The ARG cassettes for *A. baumannii* and *E. coli* were larger and more stable than the cassettes in *S. aureus*, which is consistent with a previously published observation (Chng et al., 2020). In the case of *A. baumannii*, we identified two ARG cassettes, with one containing the genes *mph(E)*, *msr(E)*, *armA*, *aadA1*, *sul1*, *aac(6′)-Ib,* and *catB8*. The ARG cassette which contained *mph(E)*, *msr(E),* and *armA* was described previously (Chng et al., 2020). For *E. coli*, the program identified a stable small cassette that shows overlap with that of *A. baumannii*, including *sul2*, *aph(3″)-Ib* and *aph(6)-Id*. When genomes of *S. aureus* were analyzed, the program BacAnt found two ARG cassettes; the first one contained the genes *bla*PC1 and *bla*R1, while the second one encoded for *mecR1* and erm*(A)*.

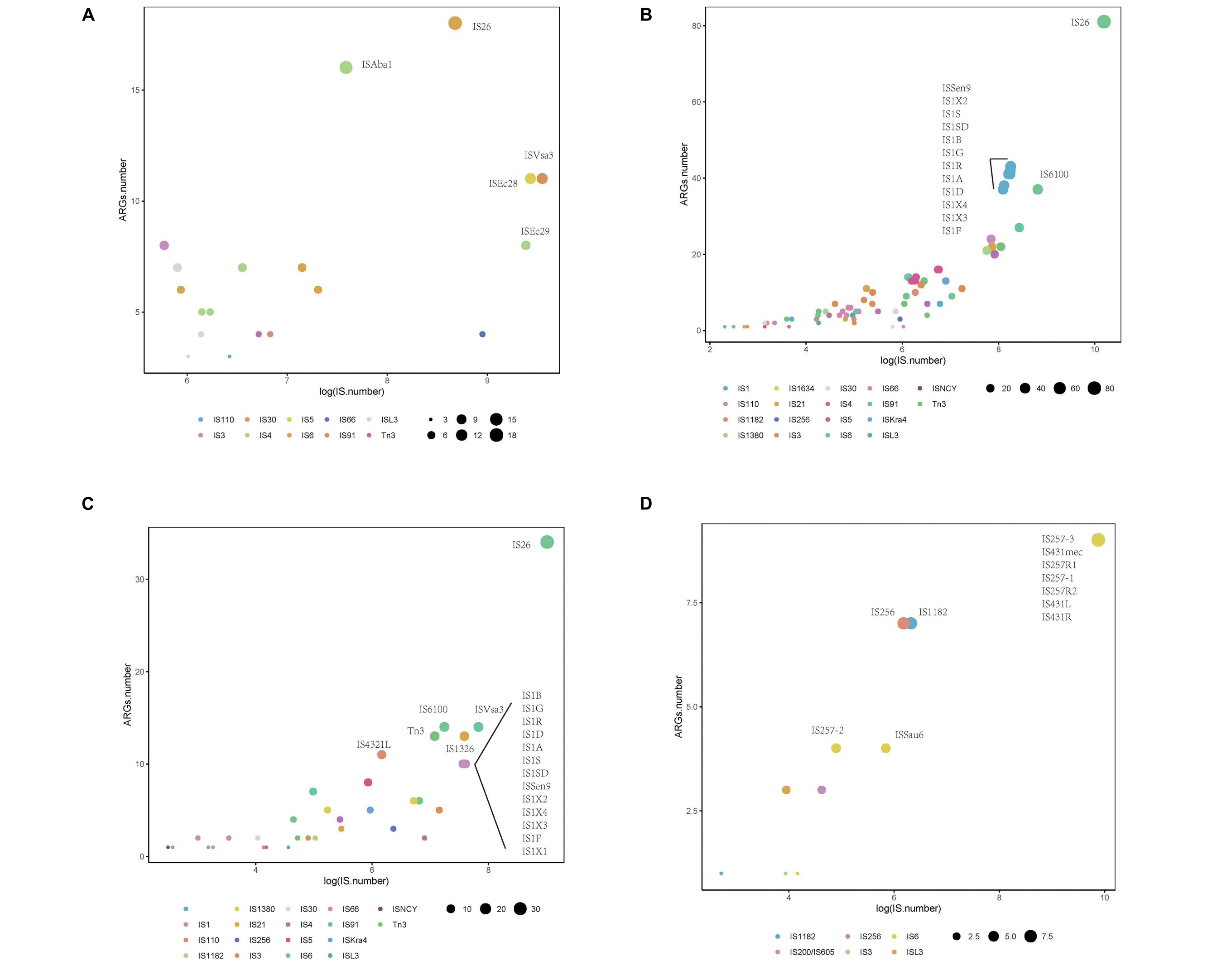


FIGURE 5 | Distribution of Insertion sequences (ISs) with statistically significant association with different types of antibiotic resistance genes (ARGs) within a 10 kb distance. Different colors show various IS, and the size of the circles indicate the presence of their associations with ARGs in **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica,* and **(D)** *S. aureus*.

We also identified commonly found ARG-IS pairs that were in close proximity to ISs (<10 kb apart). For the ARG-IS pairs, IS*Vsa3* containing the genes *aph(6)-Id*, *aph(3'')-Ib* and *tet(B)* comprised the top three ARG-IS pairs in *A. baumannii* (**Figure 4A**). For ARGs number, IS*26*, IS*Aba1,* and IS*Vsa3* were the top three active ISs (**Figure 5A**). IS*26* was the most abundant IS in *E. coli* and *S. enterica* (**Figures 4B,C**, **5B,C**). In *S. aureus*, *mecA* with diverse IS (including IS*257-3*, IS*431mec*, IS*257R1*, IS*257-1*, IS*257R2*, IS*431L,* and IS*431R*) are the top seven ARG-IS pairs (**Figures 4D**, **5D**). The result of this study confirmed that IS*6* family elements IS*26* and IS*257* play an important role in the dissemination of ARGs in *A. baumannii*, *E. coli*, *S. enterica,* and *S. aureus* (Partridge et al., 2018). As previously reported, we also observed notable differences between important MGEs in *A. baumannii*, *E. coli*, *S. enterica,* and *S. aureus* (Partridge et al., 2018).

We also analyzed the physical organization of the ARGs-IS pairs. Although the ARGs are not part of the IS, we observed a correlation of the distances between both elements in the analyzed bacterial genomes. ARGs-IS pairs occupied specific distances in *A. baumannii*, with the exception of *sul1* in IS*26* which displayed several, and more broader distance distributions (**Figure 6A**). In contrast, the *sul1* gene embedded in IS*Ec29* displayed distances that were more specific. In *E. coli*, only *mph(A)* showed a narrow distance distribution, while the other ARG-IS pairs show less correlation as the distances between the elements are less defined (**Figure 6B**). In *S. enterica*, only *aph(3'')-Ib* and IS*Vsa3*| *floR* exhibited clear positions with narrow distributions. Interestingly, the distances between IS*26* and *folR* were more widely distributed (**Figure 6C**). In *S. aureus*, *mecA*, *mecR1* and *mecI_of_mecA* exhibit specific distance distributions (**Figure 6D**). Our observations appear to indicate that the distances of ARG-IS pairs appear to often be specific, and this correlation does not appear to be defined by either the IS or the ARG alone.

Using BacAnt, we also explored the relationship between ARGs and transposons. Tn*6292* was the most commonly

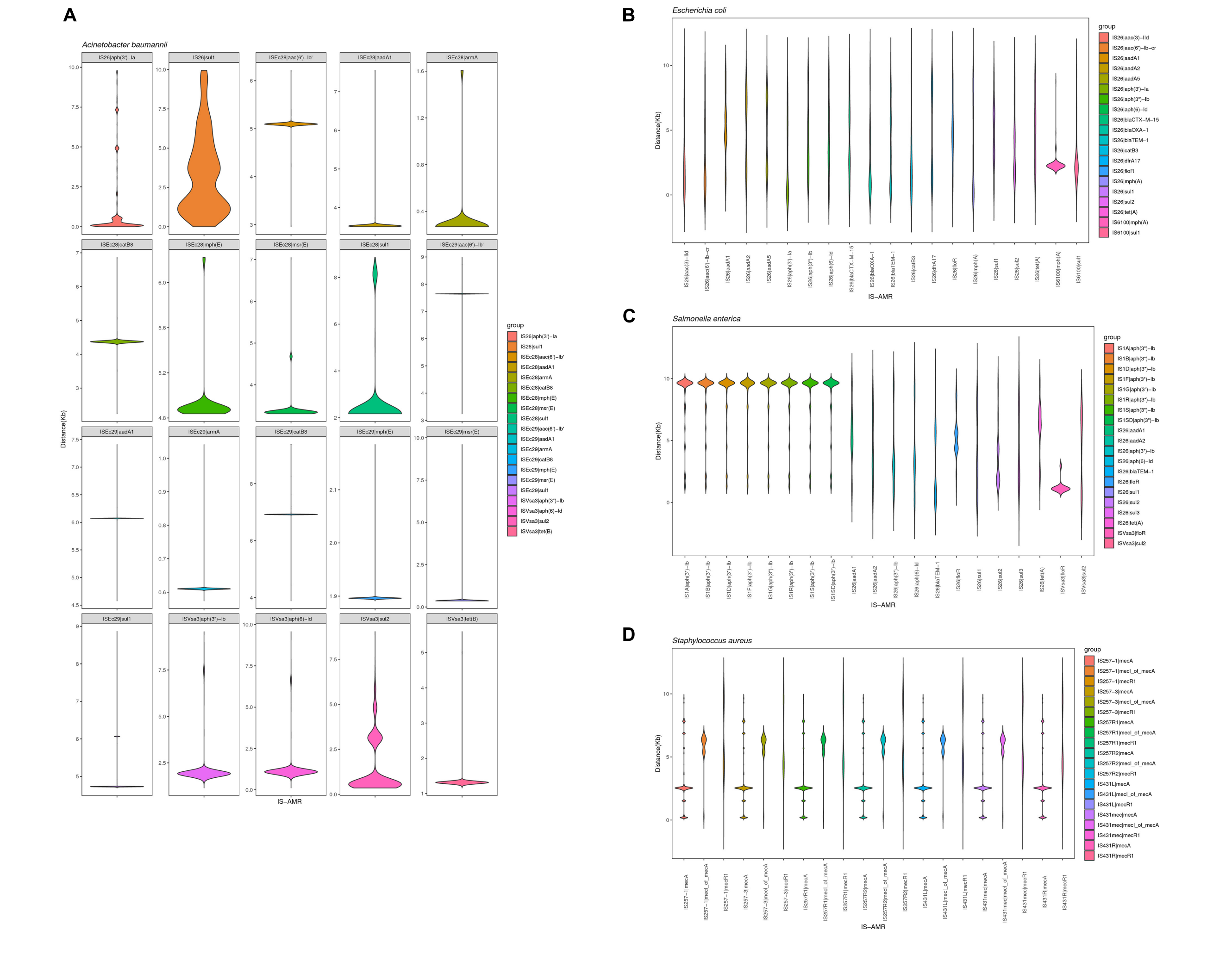


FIGURE 6 | Violin plots showing distribution of the physical distance of top 20 ARG-IS pairs in **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica,* and **(D)** *S. aureus*.

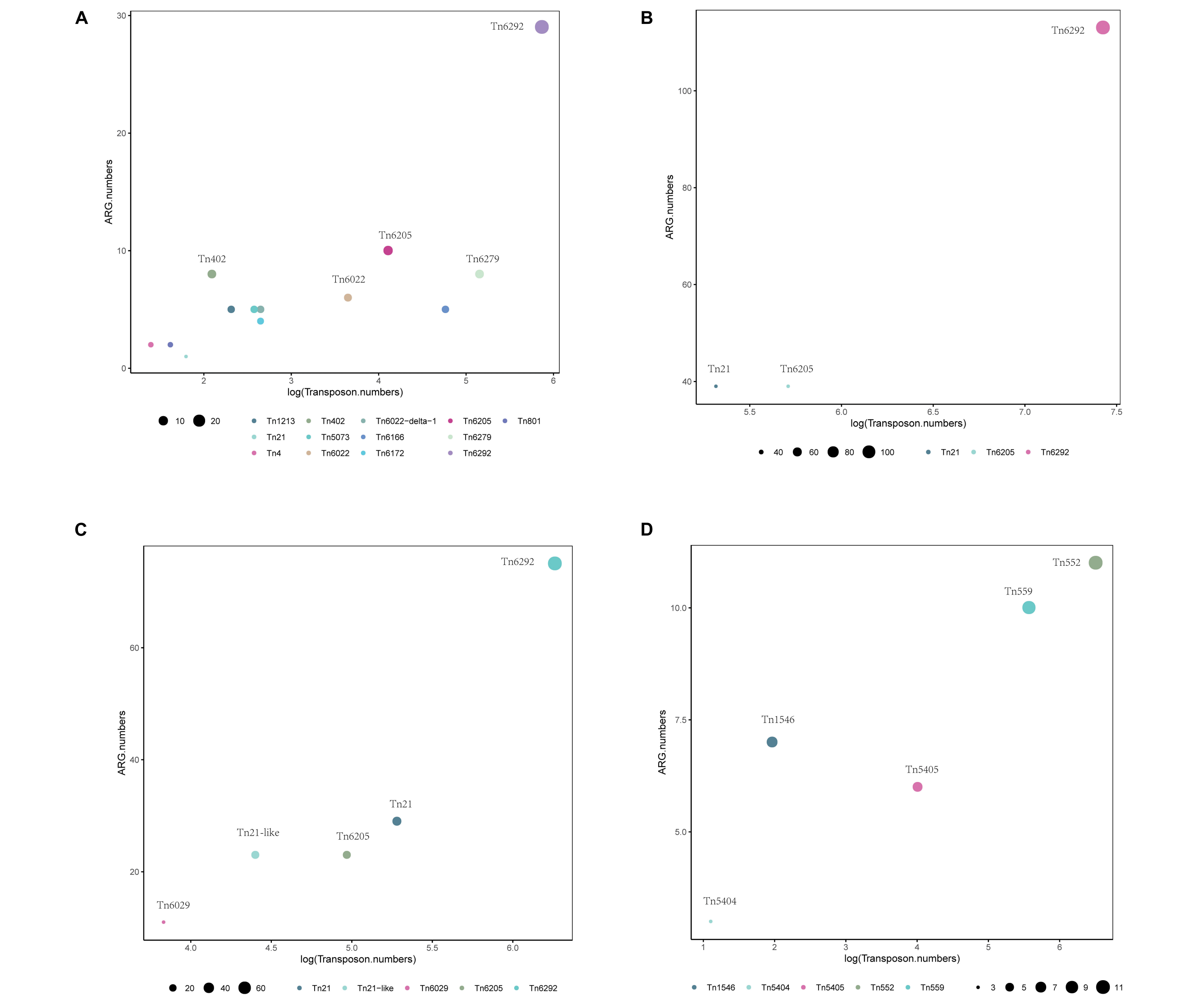


FIGURE 7 | Distribution of transposons with statistically significant association with different types of ARGs within a 10 kb distance. Different colors show various transposons, the size of the circles indicate the presence of their associated ARGs in **(A)** *A. baumannii*, **(B)** *E. coli*, **(C)** *S. enterica,* and **(D)** *S. aureus*.

observed transposon containing ARGs in *A. baumannii*, *E. coli* and *S. enterica* (**Figures 7A–C**). In *S. aureus* the most prevalent transposon with ARGs was identified to be Tn*552* (**Figure 7D**). Tn*6292* belongs to the Tn3-family and harbored an IS*26* at the right end (Chen et al., 2020). Tn*6292* also contained a quinolone resistance region *qnrS1* (Feng et al., 2016). Multidrug-resistance bacteria containing Tn*6292* are commonly observed in China (Li et al., 2018), and possibly accelerate the emergence and spread of multidrug-resistant pathogens. Tn*552* belonged to Tn*7* family, comprised of BlaZ, BlaR1, and BlaI proteins. BlaR1 is the sensor protein for the extracellular β-lactam antibiotics. The overproduction of the beta-lactamase BlaZ were responsible of β-lactam resistance. Tn*552*-like element was thought as the origin of the all β-lactamase genes in staphylococci (Gregory et al., 1997).

In order to be able to extract the maximum amount of information from whole genome sequence data, we need the improve annotation and analysis methods for MGEs (Partridge et al., 2018). In this work, we created a webserver that is easy to use and allows the annotation of ARGs, integron, and transposable elements at the same time. The pipeline generates genbank files automatically, which are compatible with easyfig for comparative genomic analysis,which will accelerate the bioinformatics analysis of ARG-related sequences.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/ **Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

HC, YY, and XH designed the study. QL, MW, and WH established the BacAnt. XH, QL, MD, WH, JW, and BL analyzed the bioinformatics data. XH, JH, WH, and SL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the grants from the National Natural Science Foundation of China (31970128, 31770142, 81861138054, and 31670135), the National Key Research and Development Project (2018YFE0101800), the Zhejiang Province Medical Platform (2020RC075 and 2017KY571), and 2019 Jiaxing Key Discipline of Medicine–infection disease (supporting subject, no. 2019-zc-02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.649969/full#supplementary-material

REFERENCES

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

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

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

Che, Y., Yang, Y., Xu, X., Brinda, K., Polz, M. F., Hanage, W. P., et al. (2021). Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2008731118. doi: 10.1073/pnas.2008731118

Chen, Q., Lin, Y., Li, Z., Lu, L., Li, P., Wang, K., et al. (2020). Characterization of a new transposon, Tn6696, on a bla NDM- 1-carrying plasmid from multidrug-resistant *Enterobacter cloacae* ssp. dissolvens in China. *Front. Microbiol.* 11:525479. doi: 10.3389/fmicb.2020.525479

Chng, K. R., Li, C., Bertrand, D., Ng, A. H. Q., Kwah, J. S., Low, H. M., et al. (2020). Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat. Med.* 26, 941–951. doi: 10.1038/s41591-020-0894-4

Cury, J., Jove, T., Touchon, M., Neron, B., and Rocha, E. P. et al. (2016). Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* 44, 4539-4550. doi: 10.1093/nar/gkw319

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., and Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* 13, 601–612. doi: 10.1038/nrg3226

Feng, W., Zhou, D., Wang, Q., Luo, W., Zhang, D., Sun, Q., et al. (2016). Dissemination of IMP-4-encoding pIMP-HZ1-related plasmids among *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* in a Chinese teaching hospital. *Sci. Rep.* 6:33419. doi: 10.1038/srep33419

Gregory, P. D., Lewis, R. A., Curnock, S. P., and Dyke, K. G. (1997). Studies of the repressor (BlaI) of beta-lactamase synthesis in *Staphylococcus aureus*. *Mol. Microbiol.* 24, 1025–1037. doi: 10.1046/j.1365-2958.1997.4051770.x

Leal, N. C., Campos, T. L., Rezende, A. M., Docena, C., Mendes-Marques, C. L., de Sa Cavalcanti, F. L., et al. (2020). Comparative genomics of *Acinetobacter baumannii* clinical strains from Brazil reveals polyclonal dissemination and selective exchange of mobile genetic elements associated with resistance genes. *Front. Microbiol.* 11:1176. doi: 10.3389/fmicb.2020.01176

Li, B., Feng, J., Zhan, Z., Yin, Z., Jiang, Q., Wei, P., et al. (2018). Dissemination of KPC-2-encoding IncX6 plasmids among multiple *Enterobacteriaceae* species in a single Chinese hospital. *Front. Microbiol.* 9:478. doi: 10.3389/fmicb.2018.00478

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi: 10.1038/nature03959

Mitchell, S. L., and Simner, P. J. (2019). Next-generation sequencing in clinical microbiology: are we there yet? *Clin. Lab. Med.* 39, 405–418. doi: 10.1016/j.cll.2019.05.003

Moura, A., Soares, M., Pereira, C., Leitao, N., Henriques, I., and Correia, A. (2009). INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* 25, 1096–1098. doi: 10.1093/bioinformatics/btp105

Partridge, S. R., Kwong, S. M., Firth, N., and Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* 31:e00088-17. doi: 10.1128/CMR.00088-17

Razavi, M., Kristiansson, E., Flach, C. F., and Larsson, D. G. J. (2020). The association between insertion sequences and antibiotic resistance genes. *mSphere* 5, e418–e420. doi: 10.1128/mSphere.00418-20

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

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

Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74, 434–452. doi: 10.1128/MMBR.00020-10

Subedi, D., Vijay, A. K., and Willcox, M. (2018). Overview of mechanisms of antibiotic resistance in *Pseudomonas aeruginosa*: an ocular perspective. *Clin. Exp. Optom.* 101, 162–171. doi: 10.1111/cxo.12621

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

Tansirichaiya, S., Rahman, M. A., and Roberts, A. P. (2019). The transposon registry. *Mob. DNA* 10:40. doi: 10.1186/s13100-019-0182-3

Wattam, A. R., Brettin, T., Davis, J. J., Gerdes, S., Kenyon, R., Machi, D., et al. (2018). Assembly, annotation, and comparative genomics in PATRIC, the all bacterial bioinformatics resource center. *Methods Mol. Biol.* 1704, 79–101. doi: 10.1007/978-1-4939-7463-4_4

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi: 10.1093/jac/dks261

Zhou, K., Lokate, M., Deurenberg, R. H., Arends, J., Lo-Ten Foe, J., Grundmann, H., et al. (2015). Characterization of a CTX-M-15 producing *Klebsiella*

Pneumoniae outbreak strain assigned to a novel sequence type (1427). *Front. Microbiol.* 6:1250. doi: 10.3389/fmicb.2015.01250

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

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fmicb.2021.687259

Whole-Genome-Based *Helicobacter pylori* Geographic Surveillance: A Visualized and Expandable Webtool

Xiaosen Jiang[1,2,3†], Zheng Xu[1,4†], Tongda Zhang[1], Yuan Li[1], Wei Li[1,2,3] and Hongdong Tan[1]*

[1] BGI-Shenzhen, Shenzhen, China, [2] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [3] College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [4] Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen, China

Helicobacter pylori exhibit specific geographic distributions that are related to clinical outcomes. Despite the high infection rate of *H. pylori* throughout the world, the genetic epidemiology surveillance of *H. pylori* still needs to be improved. This study used the single nucleotide polymorphisms (SNPs) profiling approach based on whole genome sequencing (WGS) to facilitate genomic population analyses of *H. pylori* and encourage the dissemination of microbial genotyping strategies worldwide. A total number of 1,211 public *H. pylori* genomes were downloaded and used to construct the typing tool, named *Hp*TT (*H. pylori* Typing Tool). Combined with the metadata, we developed two levels of genomic typing, including a continent-scale and a country scale that nested in the continent scale. Results showed that Asia was the largest isolate source in our dataset, while isolates from Europe and Oceania were comparatively more widespread. More specifically, Switzerland and Australia are the main sources of widespread isolates in their corresponding continents. To integrate all the typing information and enable researchers to compare their dataset against the existing global database easily and rapidly, a user-friendly website (https://db.cngb.org/HPTT/) was developed with both genomic typing tools and visualization tools. To further confirm the validity of the website, ten newly assembled genomes were downloaded and tested precisely located on the branch as we expected. In summary, the *H. pylori* typing tool (*Hp*TT) is a novel genomic epidemiological tool that can achieve high-resolution analysis of genomic typing and visualizing simultaneously, providing insights into the genetic population structure, evolution analysis, and epidemiological surveillance of *H. pylori*.

Keywords: *Helicobacter pylori*, genomic, antibiotic-resistant, phylogenetic, webtool, whole-genome sequencing, genotyping

OPEN ACCESS

Edited by:
*Liang (Leon) Wang,
Xuzhou Medical University, China*

Reviewed by:
*Yuehua Gong,
The First Affiliated Hospital of China Medical University, China
Abbas Yadegar,
Shahid Beheshti University of Medical Sciences, Iran*

****Correspondence:***
*Hongdong Tan
rtan@mgi-tech.com*

†These authors have contributed equally to this work

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

***Received:** 29 March 2021*
***Accepted:** 07 June 2021*
***Published:** 02 August 2021*

Citation:
Jiang X, Xu Z, Zhang T, Li Y, Li W and Tan H (2021) Whole-Genome-Based Helicobacter pylori Geographic Surveillance: A Visualized and Expandable Webtool. Front. Microbiol. 12:687259. doi: 10.3389/fmicb.2021.687259

INTRODUCTION

Helicobacter pylori are one of the most sophisticated colonizers in the world that infects more than half of the world's population, ranging from infants to the elderly (Suerbaum and Michetti, 2002). It is a Gram-negative bacterium that normally colonizes the gastric mucosa of humans with about 10–20% infection result in diseases (Pohl et al., 2019; Attila et al., 2020). The typical diseases that have been reported include gastritis, peptic ulcer, mucosa-associated lymphoid tissue (MALT) lymphoma, and gastric cancer (Ernst and Gold, 2000). Globally speaking, the risks of disease and the incidence and mortality of gastric cancer were geographically different (Kodaman et al., 2014).

H. pylori display a distinguished mutation rate among bacterial pathogens due to the lack of genes that initiate classical methyl-directed mismatch repair (MMR) (Alm et al., 1999). The high mutation and recombination rate made *H. pylori* genomes have enormous plasticity, facilitating this pathogen and enabling it to perfectly adapt to its host (Kang and Blaser, 2006; Didelot et al., 2013). It has been reported that *H. pylori* in chronic infection could take place through vertical and familial transmission (Schwarz et al., 2008; Ailloud et al., 2019). In within-host evolution, the mutation rate could reach ∼30 single nucleotide polymorphisms (SNPs) per genome per year (Kennemann et al., 2011), compared to *Escherichia coli* at ∼1 SNP per genome per year (Reeves et al., 2011). Taking into account this occurrence and large recombination events, a simple and efficient way to define the geographical pattern and epidemiological surveillance of *H. pylori* is needed (Yamaoka, 2009; Jolley et al., 2018).

The transmission of *H. pylori* transmission is slow, taking place mostly within a household it does not tend to spread like a rapid epidemic (Didelot et al., 2013). Their phylogeny was based on MLST genes and later whole genomes revealed a population structure primarily reflecting early human migration events especially out of Africa 60,000 years ago but not recent spreading (Falush et al., 2003). The global population was split into hp groups, each of which is split into hsp subgroups in the agreed convention. The hpEastAsia includes hspEastAsia, hspMaori, and hspAmerind (Kawai et al., 2011; Montano et al., 2015; Thorell et al., 2017).

To describe the population structure of *H. pylori*, genetic typing methods such as single gene typing (e.g., *cagA*, *vacA*) were recorded in previous studies (Salama et al., 2007; Yamaoka, 2009), while seven-gene multi-locus sequence typing (MLST) became the dominant tool in the later stage due to its simple and rapid typing strategy, which covers genes including *atpA*, *efp*, *mutY*, *ppa*, *trpC*, *urel*, *yphC* that categorize *H. pylori* into different sequence types (STs) (Achtman et al., 1999). However, the resolution of seven-gene MLST was still low, which limited us to tracing the epidemiological origins of *H. pylori* strains (Banerji et al., 2020). Comparatively, SNP typing covers comprehensive core genes that can generate a matrix comprising concatenated SNPs and location information in the genome, which facilitated the newly sequenced genomes to be comparable by mapping and increase the typing resolution.

It has been found that 7-gene MLST are also linked to regional epidemics across the world. The 7-gene MLST typing method enables the regional specific recognition based on the defined STs, in which geographical pattern is linked with the different risks of clinical disease. For example, non-African and African lineage could be associated with different risks of gastric disease (Campbell et al., 2001). Thus, geographic patterns can somehow link to the possibility of clinical disease. However, the seven-gene genotypes of *H. pylori* are diverse due to the high variability of *H. pylori* genomes, which hinders the recognition of patterns directly from the sequence types (STs) in 7-gene MLST. In addition, there is no information on geographical patterns or visualization tools for seven-gene MLST, thus such related geographic patterns were hard to find when a new ST was found.

This study describes a *H. pylori* genomic typing tool, *Hp*TT (*H. pylori* Typing Tool) that uses the SNP profiling based on whole-genome sequencing data. In addition to genomic typing, *Hp*TT also provides a phylogenetic and geographic visualization tool based on the Nextstrain framework (Hadfield et al., 2018). This tool allows users to upload *H. pylori* WGS data for genomic typing and uncover possible transmission events of *H. pylori*. It is believed that this tool can not only improve genome typing resolutions but may also predict the possible origin of the epidemic *H. pylori* isolates, enabling the global surveillance of *H. pylori*.

MATERIALS AND METHODS

Helicobacter pylori Genomes Downloaded and Filtered in This Study

A total number of 1,654 assembled *H. pylori* genomes were downloaded from the NCBI RefSeq database (genomes available as of May 4, 2020) using the ncbi-genome-download tool (version 0.2.12). The corresponding metadata of assembled genomes was searched by function using Entrez Direct (version 10.9) (Kans, 2020). By metadata filtering, 1,211 genomes were selected with sample collection location available (**Table 1**). All genomes were scanned by mlst (version 2.11) with the library of MLST updated on December 31, 2020 (Jolley and Maiden, 2010).

SNP Analysis

The 1,211 assembled genomes were mapped to the reference genome *H. pylori* 26695 (GenBank: AE000511.1) (Tomb et al., 1997) using MUMmer (version 3.23) (Kurtz et al., 2004). SNPs were filtered with a minimum mapping quality cutoff at 0.90 across 1,211 assembled *H. pylori* genomes. 6,129 SNPs were found, and an SNP profile of *H. pylori* is established for the corresponding isolates.

Phylogenetic Analysis

The maximum likelihood (ML) phylogenetic tree was constructed by iqtree (version 2.0.3) (Nguyen et al., 2015) based on 6,129 SNPs alignments of all 1,211 isolates. The reference genome *H. pylori* 26695 was used as an outgroup. The tree was generalized by the Gamma distribution to model site-specific rate variation (the GTR model). Bootstrap pseudo-analyses of the alignment were set at > = 1000. All ML trees were visualized and annotated using Figtree (version 1.4.4). The minimum spanning tree was constructed by the GrapeTree (v1.5.0) (Zhou et al., 2018). The mutation rate of the *cagA* gene was calculated by BEAST v1.8.4 (Suchard et al., 2018).

Geographic Typing System

Based on the phylogenetic tree, two levels of the geographic group were defined, including the first level defined at the continent

TABLE 1 | Summary of 1,211 *H. pylor*i genomes.

Continent	Country (region) of origin	Number of isolates
Asia		312 (25.76%)
	Cambodia	53
	China	74
	China (Taiwan)	8
	India	47
	Indonesia	1
	Japan	31
	Kuwait	2
	Malaysia	79
	North Korea	1
	Singapore	14
	South Korea	1
	Vietnam	1
Africa		10 (0.82%)
	Morocco	6
	Nigeria	1
	South Africa	3
Europe		294 (24.28%)
	Belarus	2
	Belgium	6
	France	37
	Germany	31
	Ireland	1
	Poland	2
	Portugal	1
	Russia	3
	Spain	54
	Sweden	19
	Switzerland	130
	United Kingdom	8
Oceania		178 (14.70%)
	Australia	177
	Papua New Guinea	1
North America		233 (19.24%)
	Canada	2
	El Salvador	1
	Mexico	118
	Nicaragua	24
	United States of America	88
South America		184 (15.19%)
	Angola	1
	Colombia	172
	Peru	11

scale and the second level defined as a country-specific scale. In the first level of genotyping, lineages carrying more than seven isolates and $>75\%$ isolates sourced from one major continent were defined as a continent-specific group or clade. A mixed continent group was defined when there was no major continent identified with isolates at $>75\%$. In the second level, lineages carrying more than one isolate and $>75\%$ isolates sourced from one major country were defined as a country-specific group or subclade. In addition, a mixed group was also defined at level two when there were more than two isolates and not a major country identified with isolates at $>75\%$. The association of the genomic lineage of *H. pylori* with the geographic information of isolates provided a map that allows us to trace both the possible transmission and evolution of a detected or sequenced *H. pylori* genome.

Establishment of *Helicobacter pylori* Database

The *Hp*TT website was established based on two modules: (1) The genomic-geographical typing tool of *H. pylori* isolates and (2) a visualization tool of both the genomic and geographic typing results. The online typing tool was written in PHP, Javascript, css, and HTML. The online visualization service was performed based on the CodeIgniter framework[1], tree visualization was analyzed by the augur[2] bioinformatics tool and the auspice[3] visualization tool imbedded in the Nextstrain (Hadfield et al., 2018) open source project. The *H. pylori* database was stored in a Mysql database.

RESULTS

Definition of Two Levels of Geographic Genotypes for *Helicobacter pylori*

A total of 1,211 assembled genomes with available geographic information from the NCBI RefSeq database were downloaded and analyzed for establishing the *H. pylori* genotyping database (**Supplementary Table 1**). All assembly genomes were mapped to the reference genome *H. pylori* 26695. Based on the maximum likelihood tree, 6,129 SNPs extracted from 1,135 genes on the reference genome were defined for further genomic typing. In terms of geographic information, 1,112 isolates were grouped at two levels, including 37 continent-level groups (**Figures 1A,B**) and/or 236 country-level groups (**Figures 1C,D**). The median pairwise distances (the median number of SNPs shared by the branches) between isolates were found as follows: 319 SNPs within continent clades and 1,493 SNPs within country subclades. We labeled these continent clades and country subclades using a structured hierarchical nomenclature system similar to that used for *M. tuberculosis* (Coll et al., 2014). For instance, region 1 clade (G1) is subdivided into country subclades G1.C1 and G1.C2. The mutation rate of *cagA* was 2.413×10^{-2} (95% CI: 1.600×10^{-2}–3.900×10^{-2}), which was 1.739×10^{-2}/site/year (95% CI: 1.153×10^{-2}–2.811×10^{-2}).

A Continent Level Genomic Typing for *Helicobacter pylori*

A total number of 37 continent level groups (n = 1,112) were defined, including 25 continent-specific groups and 12 mixed-continent groups (**Figures 1A,B**). Isolates across the tree did not fall into the continent group but can be defined as a country group

[1] https://www.codeigniter.com/

[2] https://github.com/nextstrain/augur

[3] https://github.com/nextstrain/auspice

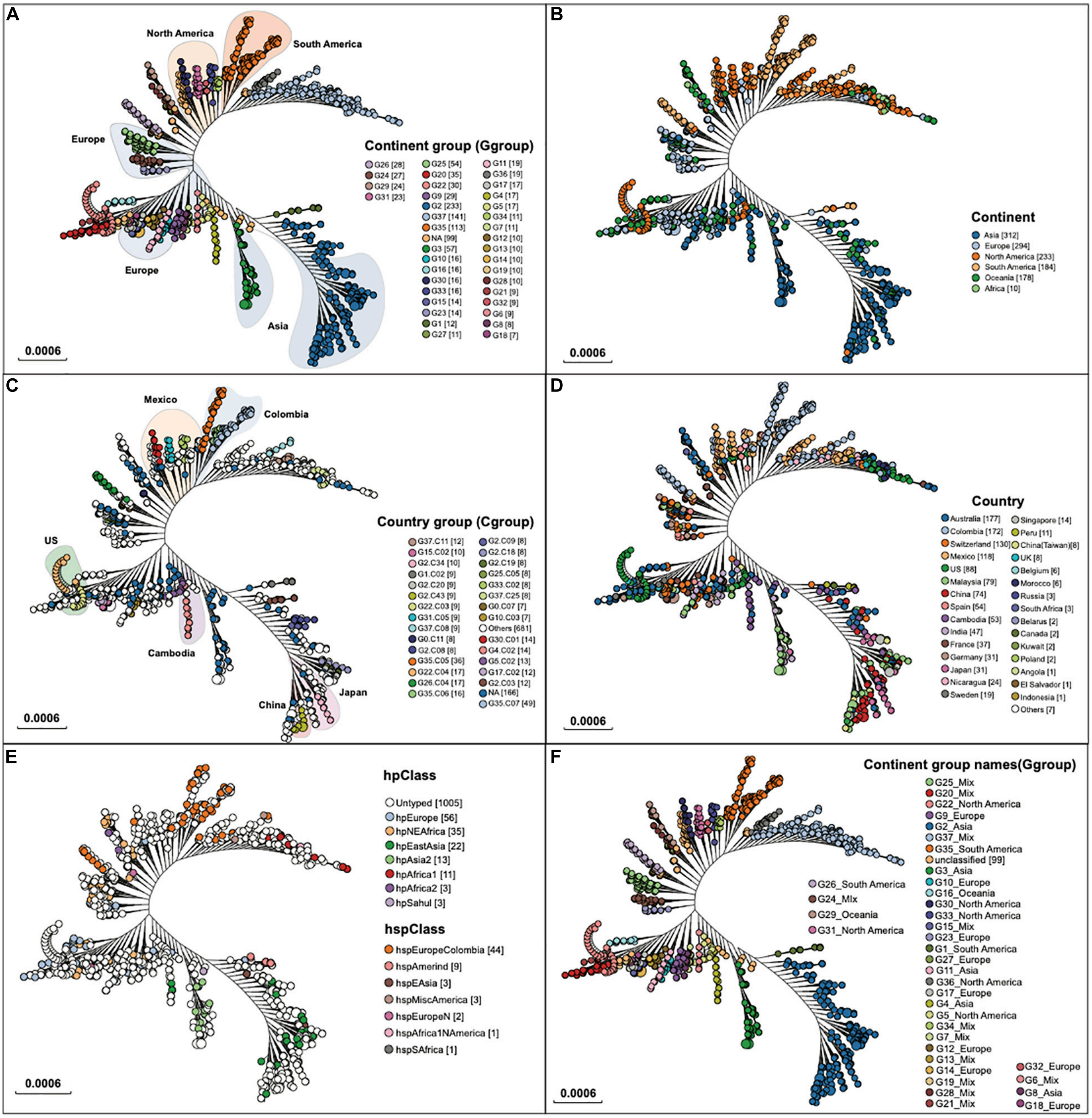


FIGURE 1 | Two clades of geographic typing based on the WGS. The *Hp*TT enrolled 1,211 *H. pylori* genomes downloaded from NCBI. The clade nodes in each figure correspond to **(A)** G groups for continent level of typing, **(B)** the continent that isolates collected from, **(C)** C groups for country-level typing, **(D)** the country that isolates were collected from. **(E)** the hp Class and hsp Class, **(F)** G groups for continent level of typing with group names. Numbers in parenthesis refer to the number of isolates in each genogroup.

that was named G0 ($n = 74$). Isolates across the tree that fell into neither fall into the country group nor the continent group were defined as non-grouped ($n = 25$). Because the genome data of *H. pylori* were downloaded from the NCBI database, and these genomes came from various regions of the world. Compared with their ancestors, these strains have different genomes, which has led to the formation of independent evolutionary branches. After they formed independent evolutionary branches, (1) they may not have spread. (2) After the spread, it was not collected. These two reasons could account for an insufficient number of strains in the branch, which cannot form a group with regional characteristics under our typing method.

Five continent-specific groups contain more than 75% Asian isolates, supporting Asia to be the continent with the largest isolate source (*n* = 319, 26.34%) (**Figure 2**). North America was found to be the second-largest group of isolate pool which consisted of six continent-specific groups (*n* = 132, 10.90%). Although fewer isolates were found to be sourced from Europe (*n* = 109, 9.00%), these isolates were distributed in nine continent-specific groups. Two groups (G16 & G29) of isolates were found to be part of the Oceania specific group (*n* = 39, 3.22%) and three groups (G1 & G26 & G35) were found to be from the South America specific groups (*n* = 109, 9.00%). In addition, the 12 mixed groups of isolates contained 226 isolates (18.66%). Among all G level groups, G2 was the largest continent specific group (*n* = 223) that mainly contained isolates from Asia (193/223, 82.83%), while G35 was the second largest continent specific group (*n* = 109) that mainly contained isolates from South America (99/109, 87.61%). Apart from all the continent groups above, there was no Africa-specific group found, but only with isolates collected from Africa defined in G28 (*n* = 2), G37 (*n* = 7), and G29 (*n* = 1) (**Figure 2**).

Although the continent-specific groups did not 100% stick to one continent in our typing system, the transmission events were still possible to predict. While most of the Asian isolates fell into the Asia groups, a small proportion of the Asian isolates belonged to the mixed groups. Similarly, most of the isolates sourced from North America and South America fell in their own region groups, while a minority of the isolates were in the mixed groups. Interestingly, isolates from Oceania and Europe could be found across all 12 mixed continent groups, reflecting the fact that *H. pylori* isolates from these two continents were relatively widespread across the globe.

The Nested Country Level Genomic Typing for *Helicobacter pylori*

A total number of 859 isolates were grouped into 216 geographic patterns at a country level, which were predominant in 29 countries across six continents (**Figure 3**). Among these 29 countries encompassing 216 groups, 20 countries found in 168 groups were defined as country-specific groups, while the remaining 9 countries were scattered over the 48 country-level mixed groups that were left.

G35.C07 was the largest country-specific group that contained 49 isolates from Colombia, followed by the G35.C05 (*n* = 35) dominated in Colombia as well. These isolates from Colombia were mainly collected from the NCBI Bioproject PRJNA352848, which study contained the population structure of *H. pylori* in regional evolution in South America (Muñoz-Ramírez et al., 2017). The isolates from groups G35.C07 and G35.C05 were mainly found in Colombia, Mexico, and Spain (**Figure 3**). This result provided evidence that the *H. pylori* isolates were possibly transmitted from Spain and spread locally in South America and North America. In comparison, Australia and Switzerland were the largest countries of isolate sources with isolates scattered across more than half of the country-specific groups.

When comparing the percentage of isolates from different countries, those isolates from France, Germany, Malaysia, Nicaragua, Sweden, and the United Kingdom were found to be scattered in more than one continent group, while isolates from Cambodia, Colombia, India, Peru, Spain, and the United States were focused in one continent group when they were also found in other continent groups. More importantly, Australia and Switzerland were two countries that were mostly found to have scattered isolates in different regional specific groups.

Three clusters were observed in the percentage of different isolate sources at continent scale (G32 to G25 with red branches in **Figure 3**), consisting of groups from Europe and mixed continents. Specifically, those isolates from mixed groups were mainly sourced from European and Oceania countries, making this cluster dominated by Europe-Oceania. The second cluster was the mixed by Asian, Oceanian, European, and mixed groups (G4 to G2 with green branches in **Figure 3**) but dominated by isolates from Australia and Asian countries. Therefore, cluster two was specified as the Asian-Pacific cluster. The third cluster was formed by North American groups (G31 to G37 with purple branches in **Figure 3**), while South American branches were next to the North American cluster.

Comparing With hp and hsp Class

hp and hsp class were designed for the geographic-genetic typing of *H. pylori* (Kawai et al., 2011; Montano et al., 2015; Thorell et al., 2017; Lamichhane et al., 2020). Of 1,211 *H. pylori* genomes, 231 were found to have been typed by hp and hsp class, which were well fit to our typing groups. Specifically, hpEastAsia, hpAsia2, and hspEAsia were included in the three Asia continent groups G2, G3, and G4 (**Figures 1E,F**), while hspEuropeColombia fell in two south America groups, G26 and G35. Similarly, hpAfrica1, hspMiscAmerica, and hspAfrica1NAmerica were mapped to a mixed group G37. The comparison with hp and hsp clusters enhanced the validity of our typing method.

Comparing With Seven-Gene MLST

Seven-gene MLST was implied to get the sequence types (STs) for all 1,211 isolates. Unfortunately, due to the high mutation rate of the *H. pylori* strains, most of the seven-gene allele were only found to have high similarity instead of an accurate type, as a result, a large number of isolates (*n* = 876, 72.3%) were untyped in our dataset (**Supplementary Table 1** and **Figure 1**). Among all the countries, Australia and Switzerland were the two countries with a higher number of untyped isolates, which is probably due to the isolates being collected by those two countries having not been submitted to the pubMLST website to be typed.

A User-Friendly Typing Website

To support our *H. pylori* geographic typing tool, a user-friendly typing website was established and made available at https://db.cngb.org/HPTT/. Our *Hp*TT approach is compatible with any whole-genome sequencing (WGS) data with metadata (**Figure 4**). For the sequencing data from pure-cultured isolates, the assembled genomes can be directly submitted to our website. However, it is worth noting that sequences or assembled genomes needed to be extracted from metagenome samples before submission (Parks et al., 2017; Olekhnovich et al., 2019). Except for the sequenced genome data, the available assembled

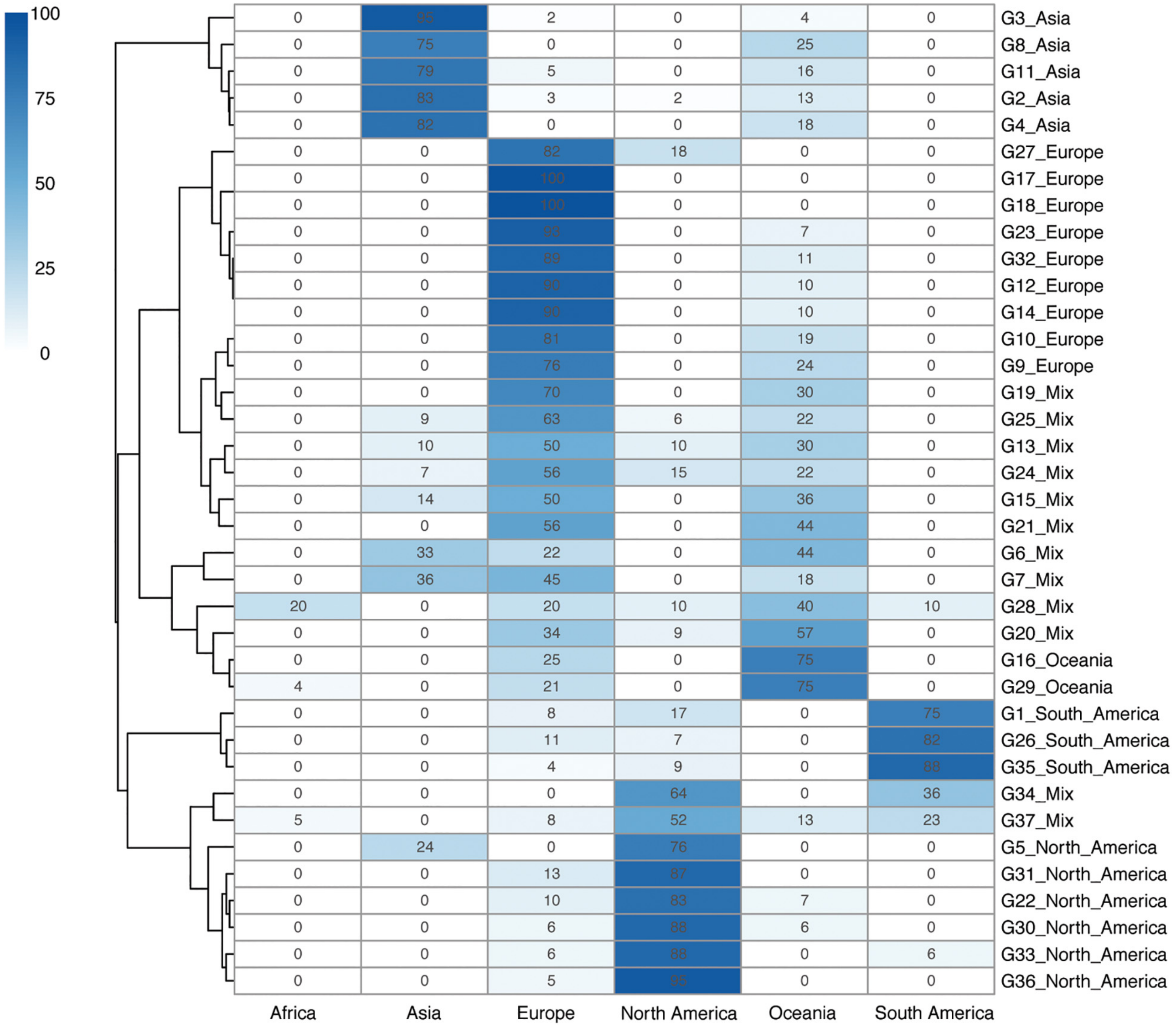


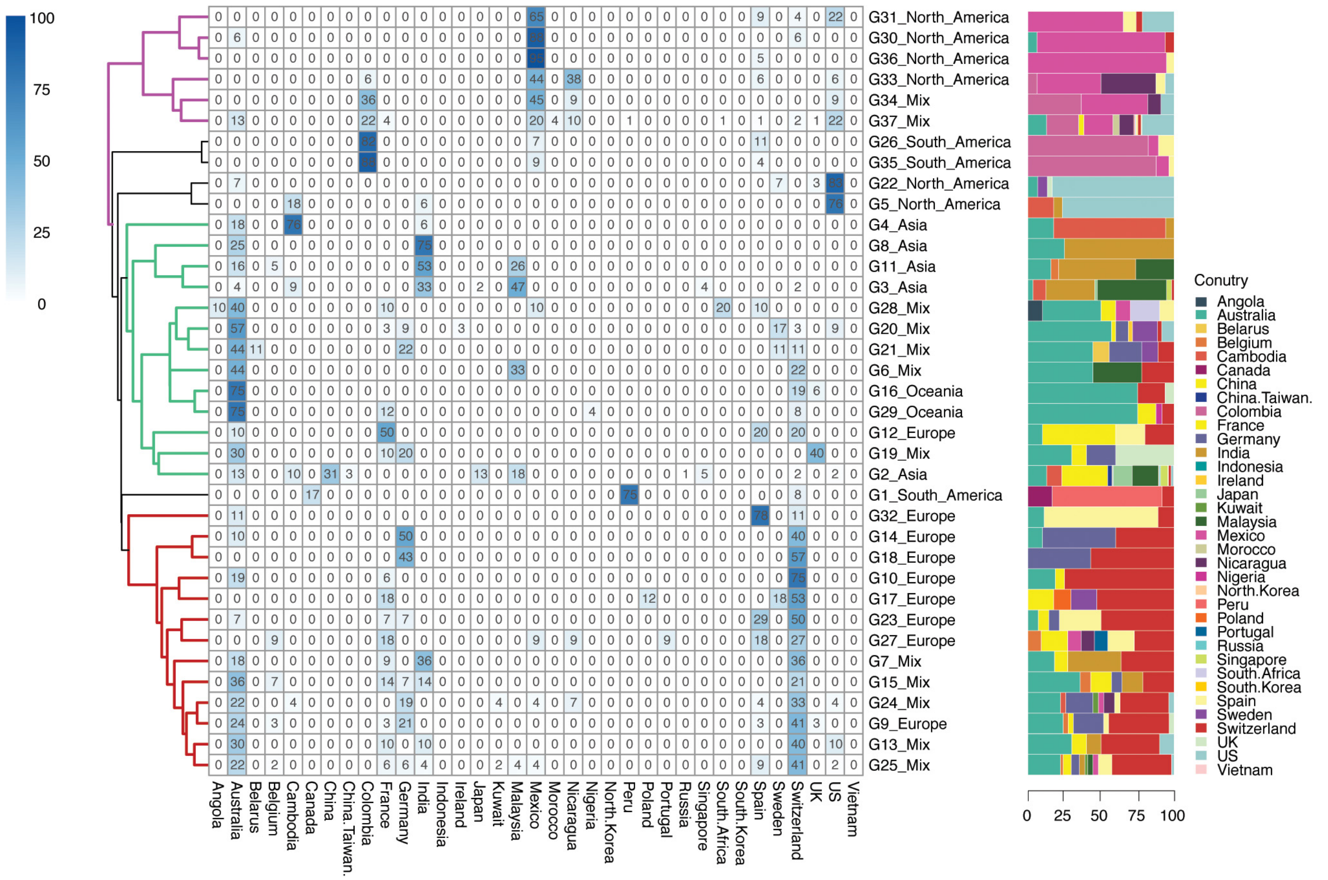
FIGURE 2 | Geographical clustering of *H. pylori* continent clades. The number in each cube represents the percentage of unique isolates sourced from each of the continents. A total number of 37 continent-level groups were defined. The deeper the color, the higher the percentage of the isolates in that continent level of clade groups. A phylogenetic tree is also shown on the left side of the table. Background information on the isolates is provided in **Supplementary Table 1**.

contigs from NCBI Sequence Read Archive (SRA) or assembly database (RefSeq), or other genome databases (e.g., European Nucleotide Achieve) can also be directly uploaded to our website. By using MUMmer alignment and blast process, the uploaded genome can be located to the closest matching genomes, further facilitating the possible transmission route analysis across the globe. In addition, our database can be also linked to the NCBI genome database, helping the user easily locate the metadata information from the available database (see **Supplementary Material**).

Except for the typing tool, the Nextstrain framework was also embedded in our website. By clicking the uploaded genome number, information can be linked to the phylogenetic tree with the corresponding continent and country. Possible evolution relationships and interactive located functions have made our typing tools easy to be applied and understood.

The Validation of Our Genomic Typing Method

For validating the accuracy of the genomic typing method and the efficiency of the web tool, ten new genomes from NCBI were downloaded and tested (**Supplementary Table 2**). Except for one genome (GCF_002206465.1), which failed due to being sequenced by Pacbio, the remaining nine genomes were typed successfully [Our typing tool was established based on the MPS (Massive Parallel Sequencing) data, Pacbio sequencing may generate many SNPs in the gap region in MPS sequencing].



	Angola	Australia	Belarus	Belgium	Cambodia	Canada	China	China.Taiwan.	Colombia	France	Germany	India	Indonesia	Ireland	Japan	Kuwait	Malaysia	Mexico	Morocco	Nicaragua	Nigeria	North.Korea	Peru	Poland	Portugal	Russia	Singapore	South.Africa	South.Korea	Spain	Sweden	Switzerland	UK	US	Vietnam
G31_North_America	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0	0	0	0	9	0	4	0	22	0
G30_North_America	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
G36_North_America	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
G33_North_America	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	44	0	38	0	0	0	0	0	0	0	0	0	6	0	0	0	6	0
G34_Mix	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	45	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0
G37_Mix	0	13	0	0	0	0	0	0	22	4	0	0	0	0	0	0	0	20	4	10	0	0	1	0	0	0	0	1	0	1	0	2	1	22	0
G26_South_America	0	0	0	0	0	0	0	0	82	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
G35_South_America	0	0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
G22_North_America	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	3	83	0
G5_North_America	0	0	0	0	18	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76	0
G4_Asia	0	18	0	0	76	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G8_Asia	0	25	0	0	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G11_Asia	0	16	0	5	0	0	0	0	0	0	0	53	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G3_Asia	0	4	0	0	9	0	0	0	0	0	0	33	0	0	2	0	47	0	0	0	0	0	0	0	0	0	4	0	0	0	0	2	0	0	0
G28_Mix	10	40	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	20	0	10	0	0	0	0	0
G20_Mix	0	57	0	0	0	0	0	0	0	3	9	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	3	0	9	0
G21_Mix	0	44	11	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	11	0	0	0
G6_Mix	0	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0
G16_Oceania	0	75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	6	0	0
G29_Oceania	0	75	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	8	0	0	0
G12_Europe	0	10	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	20	0	0	0
G19_Mix	0	30	0	0	0	0	0	0	0	10	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0
G2_Asia	0	13	0	0	10	0	31	3	0	0	0	0	0	0	13	0	18	0	0	0	0	0	0	0	0	1	5	0	0	0	0	2	0	2	0
G1_South_America	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	0	8	0	0	0
G32_Europe	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	11	0	0	0
G14_Europe	0	10	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0
G18_Europe	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57	0	0	0
G10_Europe	0	19	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0
G17_Europe	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	18	53	0	0	0
G23_Europe	0	7	0	0	0	0	0	0	0	7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	50	0	0	0
G27_Europe	0	0	0	9	0	0	0	0	0	18	0	0	0	0	0	0	0	9	0	9	0	0	0	0	9	0	0	0	0	18	0	27	0	0	0
G7_Mix	0	18	0	0	0	0	0	0	0	9	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0
G15_Mix	0	36	0	7	0	0	0	0	0	14	7	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0
G24_Mix	0	22	0	0	4	0	0	0	0	0	19	0	0	0	0	4	0	4	0	7	0	0	0	0	0	0	0	0	0	4	0	33	0	4	0
G9_Europe	0	24	0	3	0	0	0	0	0	3	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	41	3	0	0
G13_Mix	0	30	0	0	0	0	0	0	0	10	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	10	0
G25_Mix	0	22	0	2	0	0	0	0	0	6	6	4	0	0	0	2	4	4	0	0	0	0	0	0	0	0	0	0	0	9	0	41	0	2	0

FIGURE 3 | Geographical clustering of *H. pylori* country subclades. The number in each cube represents the percentage of unique isolates sourced from each of the countries in that continent group. A total number of 216 country-level groups were defined. The deeper the color, the higher the percentage of the isolates sourced from that country in continent-level groups. Background information on isolates is provided in **Supplementary Table 1**.

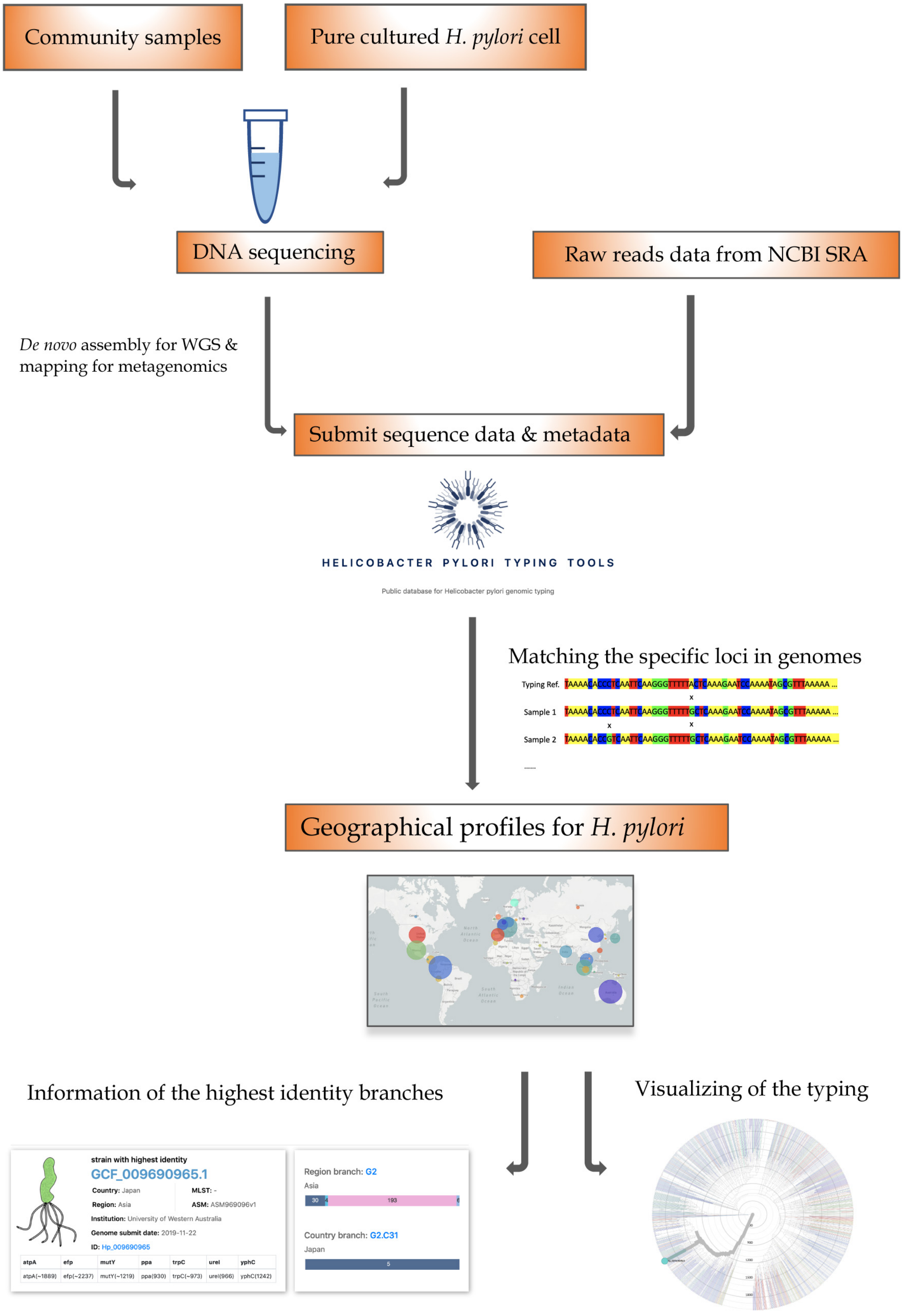
DISCUSSION

The epidemiological patterns of *H. pylori* isolates have been reported with specific geographic characteristics. In this study, the new typing webtool *Hp*TT not only illustrated the population structure of *H. pylori* but also made genomic typing easy to approach. In the continent level of typing, 1,112 isolates were grouped into 37 continent-specific patterns. Except for 12 continent mixed groups, the rest could be defined as continent-specific groups across the five continents. Isolates from Europe and Oceania were universally found in most of the continent-level groups (Europe 33/37, 89.19% and Oceania 26/37, 70.27%), illustrating that isolates from these two continents were widely spread across the world.

In the country level of typing, 1,045 isolates were grouped into 216 country-level groups. Most of the isolates were defined as country-specific groups (168/216, 77.77%), while the rest of the isolates were grouped as country mixed groups (48/216, 22.22%). Australian and Swiss isolates were found to be widespread around the world, while isolates from Columbia were more regionally specific. It has been reported that *H. pylori* in South America were originally transmitted from Spain (Muñoz-Ramírez et al., 2017), this data perfectly aligned with our results in G35.C05 and G35.C07, giving support to the accuracy of our genomic typing method.

The phylogenetic tree in this study was built by the collection of *H. pylori* genomes downloaded from the NCBI Refseq database. Ideally, all the isolates would be able to be grouped into different geographic groups, but there are still a few isolates that cannot be grouped by our typing tool due to the following reasons: (1) They have not spread after forming independent evolutionary branches, (2) After spreading, their offspring have not been collected and sequenced.

H. pylori show high and fine (~40 bp patch) intergenic recombination (Bubendorfer et al., 2016), which leads to sharing patches of genome sequences and makes the phylogenetic relationship obscure. Special methods have been developed to infer a population structure based on this sharing (Yahara et al., 2013). Although such typing methods are built based on core SNPs that cannot accurately trace the origin of the isolates comparing to a recent comprehensive study of *H. pylori* (Muñoz-Ramírez et al., 2021), we established a simple, rapid, and user-friendly genetic-geographic typing tool in the population structure description. The core SNPs of


Community samples
Pure cultured H. pylori cell
DNA sequencing
Raw reads data from NCBI SRA
De novo assembly for WGS & mapping for metagenomics
Submit sequence data & metadata
HELICOBACTER PYLORI TYPING TOOLS
Public database for Helicobacter pylori genomic typing
Matching the specific loci in genomes
Typing Ref. TAAAACACCCTCAATTCAAGGGTTTTTACTCAAAGAATCCAAAATAGCGTTTAAAAA ...
x
Sample 1 TAAAACACCCTCAATTCAAGGGTTTTTGCTCAAAGAATCCAAAATAGCGTTTAAAAA ...
x x
Sample 2 TAAAACACCGTCAATTCAAGGGTTTTTGCTCAAAGAATCCAAAATAGCGTTTAAAAA ...
......
Geographical profiles for H. pylori
Information of the highest identity branches
Visualizing of the typing
Region branch: G2
Asia
30
4
193
6
Country branch: G2.C31
Japan
5

strain with highest identity
GCF_009690965.1
Country: Japan MLST: -
Region: Asia ASM: ASM969096v1
Institution: University of Western Australia
Genome submit date: 2019-11-22
ID: Hp_009690965

atpA	efp	mutY	ppa	trpC	urel	yphC
atpA(~1889)	efp(~2237)	mutY(~1219)	ppa(930)	trpC(~973)	urel(966)	yphC(1242)

FIGURE 4 | The *Hp*TT workflow. The SNP-based genotyping approach can be used with the Whole Genome Sequencing (WGS) data, which can be acquired in the following ways: DNA can be extracted from a pure cultured bacterial cell with WGS data or a community sample with metagenomic sequencing data. After being sequenced by an appropriate platform, the assembled genomes can be directly submitted to our database. In addition, the public assembled data can also be directly submitted to our database. The downstream analyses of the aligned sequence data can be linked to the phylogenetic and geographic page.

1,211 *H. pylori* genomes were filtered with a minimum mapping quality cut off at 0.90, which means the individual indels for isolates were not kept. Our typing method has been further validated by testing genomes, suggesting that the typing tool was successfully established.

The addition of 7-gene MLST to our database intended to offer an easy way for users to visualize both results from our typing method and 7-gene MLST with comparisons. The large set of untyped isolates in 7-gene MLST might be related to the insufficient submission of genomes to pubMLST. In our typing database, isolates collected from Australia and Switzerland were scattered across different regional groups, which might be due to the frequent transmission event that occurred between Australia/Switzerland and other countries.

In this study, except for the novel typing tool, a user-friendly website was also established. By using this typing tool, users can achieve fast and precise genomic typing, easily locating the possible origins and transmission events across the world. When located in the actual geographic group, it is easy for users to check the details of the corresponding components of the branches in our database. The genome with the highest identity can be easily linked to the NCBI database as well as the visualization tool where the dynamic evolution of *H. pylori* was shown. At the same time, seven-gene MLST results were displayed for each genome in the database, as well as the hp groups and hsp subgroup results studied previously (Kawai et al., 2011; Yahara et al., 2013).

The most interesting part of the *Hp*TT tool and methodology allows us to perform genome typing with assembled genomes from the metagenomics samples, as illustrated in **Figure 4**. Due to the rapid mutation of *H. pylori*, it is most likely that the sample from one's gut is heterogeneous. Whole-genome sequencing by combining sequencing libraries labeled with different barcodes on a meta sample, and a cultured pure isolate could yield enough data from one single run to perform the epidemiological surveillance of *H. pylori* on a global level to find the possible transmission event in evolution profile. An open-source assay protocol will be developed and shared in the future to combine with this *Hp*TT tool to enable the epidemiological surveillance of *H. pylori*.

Although our typing tool filled a gap in the genetic epidemiological surveillance of *H. pylori*, some functions still need to be improved. For example, cytotoxin-associated gene A (*cagA*) and *vacA* were two crucial genes that were reported to be correlated with geographic patterns of *H. pylori* (Yamaoka, 2009; Breurec et al., 2011). The *cagA* gene is one of the most important virulence genes in *H. pylori*, located at the end of a cag pathogenicity island (cag PAI) that encodes 120–145 kDa CagA protein (Šterbenc et al., 2019). Another virulence factor was vacuolating cytotoxin encoded by the gene *vacA* (Šterbenc et al., 2019). The variation of these two genes was widely reported by the *H. pylori* groups that can reflect the genomic difference for different geographic patterns. However, such a rapid typing method on a website for these two genes is still lacking, which could be considered in the further *Hp*TT version 2.

H. pylori are normally treated by antibiotics without antimicrobial susceptibility testing (Pohl et al., 2019). Antibiotics-resistant *H. pylori* has been reported related to several mutations within the genes *pbp1A*, *23S rRNA*, *gyrA*, *rdxA*, *frxA*, and *rpoB* (Domanovich-Asor et al., 2021). These antibiotics-resistant genes will be included in the second version despite there already being an antibiotics-specific resource available (Yusibova et al., 2020). As more strains or isolates are being deposited into our database along with geographic information, *Hp*TT could be more powerfully associate genomic typing with geographic information and phenotypes.

In summary, this work illustrates efforts in a global epidemiological study of *H. pylori* isolates. Two functions were designed for the web typing tool, one for genomic typing and the other for phylogenetic and geographic visualization. The accuracy of our genomic typing system was proved by ten unused genomes as well as in another published study (Muñoz-Ramírez et al., 2017). Together with the visualization tool, the genomic population structure of *H. pylori* with geographic documents were described. Future studies will be expanded by the crucial virulence gene and antibiotic-related genes. This tool is beneficial for the surveillance of *H. pylori* for public health and the monitoring of its epidemic development.

DATA AVAILABILITY STATEMENT

All assembled *H. pylori* genomes used in this study were downloaded from NCBI assembly database (https://www.ncbi.nlm.nih.gov/assembly/) under the accession numbers in **Supplementary Tables 1, 2**.

AUTHOR CONTRIBUTIONS

ZX and HT conceived the study. XJ performed the analysis. TZ, YL and WL revised the manuscript. HT provided critical analysis and discussions. All authors discussed the results and contributed to the revision of the final manuscript.

FUNDING

This study was supported by the Science, Technology, and Innovation Commission of Shenzhen Municipality under a grant (No. JCYJ20170412153155228).

ACKNOWLEDGMENTS

We thank China National GeneBank at Shenzhen for supporting this study. We wish to thank Daoming Wang by the timely help of genome downloading and analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.687259/full#supplementary-material

REFERENCES

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z. J., et al. (1999). Recombination and clonal groupings within *Helicobacter* pylori from different geographical regions. *Mole. Microb.* 32, 459–470. doi: 10.1046/j.1365-2958.1999.01382.x

Ailloud, F., Didelot, X., Woltemate, S., Pfaffinger, G., Overmann, J., Bader, R. C., et al. (2019). Within-host evolution of *Helicobacter* pylori shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nat. Comm.* 10, 1–13.

Alm, R. A., Ling, L.-S. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., et al. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter* pylori. *Nature* 397, 176–180. doi: 10.1038/16495

Attila, T., Zeybel, M., Yigit, Y. E., Baran, B., Ahishali, E., Alper, E., et al. (2020). Upper socioeconomic status is associated with lower *Helicobacter* pylori infection rate among patients undergoing gastroscopy. *J. Infect. Dev. Count.* 14, 298–303. doi: 10.3855/jidc.11877

Banerji, S., Simon, S., Tille, A., Fruth, A., and Flieger, A. (2020). Genome-based *Salmonella* serotyping as the new gold standard. *Sci. Rep.* 10, 1–10.

Breurec, S., Guillard, B., Hem, S., Papadakos, K. S., Brisse, S., Huerre, M., et al. (2011). Expansion of European vacA and cagA alleles to East-Asian *Helicobacter* pylori strains in Cambodia. *Infect. Genet. Evol.* 11, 1899–1905. doi: 10.1016/j.meegid.2011.08.007

Bubendorfer, S., Krebes, J., Yang, I., Hage, E., Schulz, T. F., Bahlawane, C., et al. (2016). Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter* pylori. *Nat. Comm.* 7, 1–12.

Campbell, D. I., Warren, B. F., Thomas, J. E., Figura, N., Telford, J. L., and Sullivan, P. B. (2001). The African enigma: low prevalence of gastric atrophy, high prevalence of chronic inflammation in West African adults and children. *Helicobacter* 6, 263–267. doi: 10.1046/j.1083-4389.2001.00047.x

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigao, J., Viveiros, M., et al. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Comm.* 5, 1–5.

Didelot, X., Nell, S., Yang, I., Woltemate, S., Van der Merwe, S., and Suerbaum, S. (2013). Genomic evolution and transmission of *Helicobacter* pylori in two South African families. *Proc. Natl. Acad. Sci.* 110, 13880–13885. doi: 10.1073/pnas.1304681110

Domanovich-Asor, T., Craddock, H. A., Motro, Y., Khalfin, B., Peretz, A., and Moran-Gilad, J. (2021). Unraveling antimicrobial resistance in *Helicobacter* pylori: Global resistome meets global phylogeny. *Helicobacter* 2021:e12782.

Ernst, P. B., and Gold, B. D. (2000). The disease spectrum of *Helicobacter* pylori: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Ann. Rev. Microbiol.* 54, 615–640. doi: 10.1146/annurev.micro.54.1.615

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., et al. (2003). Traces of human migrations in *Helicobacter* pylori populations. *Science* 299, 1582–1585. doi: 10.1126/science.1080857

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. doi: 10.1093/bioinformatics/bty407

Jolley, K. A., Bray, J. E., and Maiden, M. C. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST. org website and their applications. *Wellcome Open Res.* 2018:3.

Jolley, K. A., and Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.* 11, 1–11. doi: 10.1186/1471-2105-11-595

Kang, J., and Blaser, M. J. (2006). Bacterial populations as perfect gases: genomic integrity and diversification tensions in *Helicobacter* pylori. *Nat. Rev. Microbiol.* 4, 826–836. doi: 10.1038/nrmicro1528

Kans, J. (2020). *Entrez direct: E-utilities on the UNIX command line in Entrez Programming Utilities Help [Internet]*. Bethesda: National Center for Biotechnology Information.

Kawai, M., Furuta, Y., Yahara, K., Tsuru, T., Oshima, K., Handa, N., et al. (2011). Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter* pylori East Asian genomes. *BMC Microbiol.* 11, 1–28.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., et al. (2011). *Helicobacter* pylori genome evolution during human infection. *Proc. Natl. Acad. Sci.* 108, 5033–5038.

Kodaman, N., Pazos, A., Schneider, B. G., Piazuelo, M. B., Mera, R., Sobota, R. S., et al. (2014). Human and *Helicobacter* pylori coevolution shapes the risk of gastric disease. *Proc. Nat. Acad. Sci.* 111, 1455–1460.

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.

Lamichhane, B., Wise, M. J., Chua, E. G., Marshall, B. J., and Tay, C. Y. (2020). A novel taxon selection method, aimed at minimizing recombination, clarifies the discovery of a new sub-population of *Helicobacter* pylori from Australia. *Evol. Appl.* 13, 278–289. doi: 10.1111/eva.12864

Montano, V., Didelot, X., Foll, M., Linz, B., Reinhardt, R., Suerbaum, S., et al. (2015). Worldwide population structure, long-term demography, and local adaptation of *Helicobacter* pylori. *Genetics* 200, 947–963. doi: 10.1534/genetics.115.176404

Muñoz-Ramírez, Z. Y., Mendez-Tenorio, A., Kato, I., Bravo, M. M., Rizzato, C., Thorell, K., et al. (2017). Whole genome sequence and phylogenetic analysis show *Helicobacter* pylori strains from Latin America have followed a unique evolution pathway. *Front. Cell. Infect. Microb.* 7:50. doi: 10.3389/fcimb.2017.00050

Muñoz-Ramírez, Z. Y., Pascoe, B., Mendez-Tenorio, A., Mourkas, E., Sandoval-Motta, S., Perez-Perez, G., et al. (2021). A 500-year tale of co-evolution, adaptation, and virulence: *Helicobacter* pylori in the Americas. *ISME J.* 15, 78–92. doi: 10.1038/s41396-020-00758-0

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

Olekhnovich, E. I., Manolov, A. I., Samoilov, A. E., Prianichnikov, N. A., Malakhova, M. V., Tyakht, A. V., et al. (2019). Shifts in the human gut microbiota structure caused by quadruple *Helicobacter* pylori eradication therapy. *Front. Microb.* 10:1902. doi: 10.3389/fmicb.2019.01902

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microb.* 2, 1533–1542. doi: 10.1038/s41564-017-0012-7

Pohl, D., Keller, P. M., Bordier, V., and Wagner, K. (2019). Review of current diagnostic methods and advances in *Helicobacter* pylori diagnostics in the era of next generation sequencing. *World J. Gastroent.* 25:4629. doi: 10.3748/wjg.v25.i32.4629

Reeves, P. R., Liu, B., Zhou, Z., Li, D., Guo, D., Ren, Y., et al. (2011). Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLoS One* 6:e26907. doi: 10.1371/journal.pone.0026907

Salama, N. R., Gonzalez-Valencia, G., Deatherage, B., Aviles-Jimenez, F., Atherton, J. C., Graham, D. Y., et al. (2007). Genetic analysis of *Helicobacter* pylori strain populations colonizing the stomach at different times postinfection. *J. Bacteriol.* 189, 3834–3845. doi: 10.1128/jb.01696-06

Schwarz, S., Morelli, G., Kusecek, B., Manica, A., Balloux, F., Owen, R. J., et al. (2008). Horizontal versus familial transmission of *Helicobacter* pylori. *PLoS Pathog.* 4:e1000180. doi: 10.1371/journal.ppat.1000180

Šterbenc, A., Jarc, E., Poljak, M., and Homan, M. (2019). *Helicobacter* pylori virulence genes. *World J Gastroenterol.* 25:4870. doi: 10.3748/wjg.v25.i33.4870

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016.

Suerbaum, S., and Michetti, P. (2002). *Helicobacter* pylori infection. *New Engl. J. Med.* 347, 1175–1186.

Thorell, K., Yahara, K., Berthenet, E., Lawson, D. J., Mikhail, J., Kato, I., et al. (2017). Rapid evolution of distinct *Helicobacter* pylori subpopulations in the Americas. *PLoS Genet.* 13:e1006546. doi: 10.1371/journal.pgen.1006546

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., et al. (1997). The complete genome sequence of the gastric pathogen *Helicobacter* pylori. *Nature* 388, 539–547.

Yahara, K., Furuta, Y., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., et al. (2013). Chromosome painting in silico in a bacterial species reveals fine population structure. *Mole. Biol. Evol.* 30, 1454–1464. doi: 10.1093/molbev/mst055

Yamaoka, Y. (2009). *Helicobacter* pylori typing as a tool for tracking human migration. *Clinical Microbiol. Infect.* 15, 829–834. doi: 10.1111/j.1469-0691.2009.02967.x

Yusibova, M., Hasman, H., Clausen, P. T. L. C., Imkamp, F., Wagner, K., and Andersen, L. P. (2020). CRHP Finder, a webtool for the detection of clarithromycin resistance in *Helicobacter* pylori from whole-genome sequencing data. *Helicobacter* 25: e12752.

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

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

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

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/fmicb.2021.712886

Prediction of Minimal Inhibitory Concentration of Meropenem Against *Klebsiella pneumoniae* Using Metagenomic Data

Rundong Tan[1,2†], Anqi Yu[1,2†], Ziming Liu[3], Ziqi Liu[4], Rongfeng Jiang[1,2], Xiaoli Wang[5], Jialin Liu[5]*, Junhui Gao[1,2]* and Xinjun Wang[6]

[1] Shanghai Biotecan Pharmaceuticals Co., Ltd., Shanghai, China, [2] Shanghai Zhangjiang Institute of Medical Innovation, Shanghai, China, [3] Medical Information Engineering, Department of Medical Information, Harbin Medical University, Harbin, China, [4] Department of Biostatistics, School of Global Public Health, New York University, New York, NY, United States, [5] Department of Critical Care Medicine, Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [6] Translational Medical Center for Stem Cell Therapy, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China

Minimal inhibitory concentration (MIC) is defined as the lowest concentration of an antimicrobial agent that can inhibit the visible growth of a particular microorganism after overnight incubation. Clinically, antibiotic doses for specific infections are determined according to the fraction of MIC. Therefore, credible assessment of MICs will provide a physician valuable information on the choice of therapeutic strategy. Early and precise usage of antibiotics is the key to an infection therapy. Compared with the traditional culture-based method, the approach of whole genome sequencing to identify MICs can shorten the experimental time, thereby improving clinical efficacy. *Klebsiella pneumoniae* is one of the most significant members of the genus *Klebsiella* in the Enterobacteriaceae family and also a common non-social pathogen. Meropenem is a broad-spectrum antibacterial agent of the carbapenem family, which can produce antibacterial effects of most Gram-positive and -negative bacteria. In this study, we used single-nucleotide polymorphism (SNP) information and nucleotide *k*-mers count based on metagenomic data to predict MICs of meropenem against *K. pneumoniae*. Then, features of 110 sequenced *K. pneumoniae* genome data were combined and modeled with XGBoost algorithm and deep neural network (DNN) algorithm to predict MICs. We first use the XGBoost classification model and the XGBoost regression model. After five runs, the average accuracy of the test set was calculated. The accuracy of using nucleotide *k*-mers to predict MICs of the XGBoost classification model and XGBoost regression model was 84.5 and 89.1%. The accuracy of SNP in predicting MIC was 80 and 81.8%, respectively. The results show that XGBoost regression is better than XGBoost classification in both nucleotide *k*-mers and SNPs to predict MICs. We further selected 40 nucleotide *k*-mers and 40 SNPs with the highest correlation with MIC values as features to retrain the XGBoost regression model and DNN regression model. After 100 and 1,000 runs, the results show that the accuracy of the two models was improved.

OPEN ACCESS

Edited by:
Liang Wang,
Institut Pasteur of Shanghai, Chinese Academy of Sciences (CAS), China

Reviewed by:
Xin Liu,
Xuzhou Medical University, China
Weidong Qian,
Shaanxi University of Science and Technology, China

****Correspondence:***
Jialin Liu
liujialin77@163.com
Junhui Gao
Jhgao68@163.com

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

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

Received: *21 May 2021*
Accepted: *26 July 2021*
Published: *23 August 2021*

Citation:
Tan R, Yu A, Liu Z, Liu Z, Jiang R, Wang X, Liu J, Gao J and Wang X (2021) Prediction of Minimal Inhibitory Concentration of Meropenem Against Klebsiella pneumoniae Using Metagenomic Data. Front. Microbiol. 12:712886. doi: 10.3389/fmicb.2021.712886

The accuracy of the XGBoost regression model for *k*-mers, SNPs, and *k*-mers & SNPs was 91.1, 85.2, and 91.3%, respectively. The accuracy of the DNN regression model was 91.9, 87.1, and 91.8%, respectively. Through external verification, some of the selected features were found to be related to drug resistance.

Keywords: *Klebsiella pneumoniae*, minimum inhibitory concentration, meropenem, XGBoost, deep neural network

INTRODUCTION

Klebsiella pneumoniae is a member of thew enterobacter *Klebsiella*; it is a Gram-negative bacterium that causes one-third of all Gram-negative infections (Navon-Venezia et al., 2017). Over the past two decades, *K. pneumoniae* has undergone complex evolution, with the emergence of many high-risk, highly infectious sequence types, resulting in the sustained global spread of *K. pneumoniae* (Navon-Venezia et al., 2017). In addition to widespread transmission, the increase in drug resistance in *K. pneumoniae* is also an important issue. Many studies and reports indicate that antimicrobial resistance (AMR) strains of *K. pneumoniae* have increased at an alarming rate in recent years (Long et al., 2017; Navon-Venezia et al., 2017).

Carbapenem antibiotics play an important role in the treatment of severe infections of drug-resistant Enterobacteriaceae, and the increase of drug resistance of *K. pneumoniae* and the emergence and spread of drug-resistant strains pose a serious threat to public health (Spagnolo et al., 2014). In fact, carbapenem antibiotic resistance in *K. pneumoniae* has emerged many years ago and has spread widely around the world (Spagnolo et al., 2014). Recent studies have shown that the resistance rates of *K. pneumoniae* to aztreonam, ceftazidime, ciprofloxacin, cefotaxime, cefepime and imipenem are more than 50% (Effah et al., 2020). Meropenem has good *in vitro* anti-*K. pneumoniae* properties and is likely to have optimal bactericidal efficacy for the treatment of *K. pneumoniae* (Baldwin et al., 2008).

Meropenem belongs to the carbapenem class of antibiotics and is one of the widely used antibiotics for the treatment of *K. pneumoniae* infections, with broad-spectrum *in vitro* resistance to both Gram-positive and Gram-negative pathogens (Navon-Venezia et al., 2017). It readily penetrates the cell walls of most Gram-negative and -positive bacteria to reach its target penicillin-binding protein (PBPS) and exhibits stability to hydrolysis by most β-lactamases, including penicillinases and cephalosporinases produced by Gram-positive and Gram-negative bacteria (Navon-Venezia et al., 2017).

In addition to the selection of antimicrobial agents, the timing and dosage of effective antimicrobial agents are also very important. In general, treatment is most effective when effective antibiotics are administered early. In a study of patients with infectious shock, there was a strong relationship between time to effective antimicrobial drug onset and in-hospital mortality (corrected ratio 1.119 per hour delay) (Pesesky et al., 2016). Neither too high nor too low a dose of antibiotics is the optimal treatment regimen: too high may result in increased resistance to *K. pneumoniae*, and too low will not achieve the desired effect of treatment with antibiotics. The minimum inhibitory concentration (MIC) indicates the appropriate dosage of antibiotics. MIC is an important index to measure both the effectiveness of antimicrobial agents and bacterial resistance to drugs.

Treatment with the optimal dose of effective antibiotics as soon as possible after the infection is the key to curing *K. pneumoniae* infection. Therefore, the time required to determine the MIC is an important factor to determine whether antibiotics can be used in the early stage of infection. There are many traditional methods of MIC determination, such as spatial gas chromatography methods for antimicrobial screening, electronic testing methods, and traditional petri dish measurement methods. However, traditional methods often take 18 to 24 h or even more. In order to meet the demand for antibiotic therapy, we need to find newer, faster, and more accurate techniques for detecting the MIC of antibiotics.

In recent years, many researchers used machine learning methods to build models that can predict MIC value more quickly and accurately (Li et al., 2016, 2017; Eyre et al., 2017; Nguyen et al., 2018; Pataki et al., 2020). These papers presented the methods and models that were used to predict the MICs of *K. pneumoniae* (Nguyen et al., 2018), antibiotic moldus of *Neisseria gonorrhoeae* (Eyre et al., 2017), *Streptococcus pneumoniae* (Li et al., 2016), non-typhoid *Salmonella* (Nguyen et al., 2019), and *Escherichia coli* (Pataki et al., 2020).

A previous study has built XGBoost machine learning models to predict MICs for a comprehensive population-based collection of clinical isolates of *K. pneumoniae*, which was able to rapidly predict MICs for 20 antibiotics with an average accuracy of 92% (Nguyen et al., 2018). According to this, our study is dedicated to constructing models that can predict MICs for Meropenem treatment of *K. pneumoniae* more accurately and analyzing features that are highly correlated with MIC prediction and externally validating these features.

In this study, we first obtained single-nucleotide polymorphism (SNP) information and nucleotide *k*-mers ($k = 6, 8, 10$) counting information based on metagenomic data of *K. pneumoniae* sequence analysis and then trained the dataset with three machine learning and deep learning methods – XGBoost classification method, XGBoost regression method, and deep neural network (DNN) regression method – and finally compare the prediction results of the three methods and select the features that are highly related to MIC to construct a new prediction model to achieve higher prediction accuracy.

MATERIALS AND METHODS

Data Collection

Two types of data were included in our study: *K. pneumoniae* metagenomic sequences, and the related MIC values of the antibiotic meropenem. The metagenomic data were pre-processed as tables of *k*-mers and SNPs for further model construction and prediction. Sequenced *K. pneumoniae* genome data used in this study can be downloaded *via* BioProject with access numbers PRJNA376414, PRJNA386693, and PRJNA396774. We collected data related to the antibiotic meropenem with complete sequence information and correct scaffold assembly, and finally, the 110 genome was involved in the study. The SRA access number for each genome is shown in the supplementary table.

HS11286[1] was selected to be our reference genome for SNP calling. The table file with SRA ID and MIC values for meropenem was downloaded from the supplementary materials attached from Nguyen et al. (2018).

For sequence data, the fastq-dump tool SRA Toolkit was used (with -I –split-files parameters). SPAdes (Bankevich et al., 2012) was then used to (with −1, −2 and -o parameters) assemble the pair the end sequence for each sample. Finally, the assembled scaffold.fasta files were mapped to the reference genome to obtain *k*-mers and SNP information.

Data Pre-processing

Nucleotide *k*-mers

In the study, 110 assembled genome scaffold files were processed to produce matrices of *k*-mers features. For each genome, we cut the scaffold sequences starting from the first nucleotide with 6-, 8-, and 10-nucleotide window lengths, respectively. For the following cuts, starting points of the windows move forward with one nucleotide each time until the sequence ends. Finally, a matrix with 110 rows and 559,494 columns of 6, 8, and 10 length nucleotide fragments were created for model training.

Calling SNPs

According to studies by Yang et al. (2018, 2019), SNPs resistant to *Mycobacterium tuberculosis* were used as features for prediction.

We extracted SNPs from the whole gene to find the resistant SNPs. For SNP calling, the raw 110 *K. pneumoniae* metagenomic samples were mapped to the HS11286 ("see text footnote 1") reference genome with single end reads mode, and then reads of the 110 genome samples were mapped to the reference genome using samtoolsv1.9 (Bonfield et al., 2021) and resulting in 110.vcf files. Further filtering was conducted using bcftools v1.10 (Li, 2011) (with parameters %QUAL $\geq$ 50 & DP $\geq$ 20). Finally, a combined matrix of the combined SNPs with 110 rows and 164,138 columns was obtained. The columns of the matrix represent the concatenation of the SNP positions compared to the reference genome, where a sample with a mutation at that position was marked as 1 and those without mutations were marked as 0.

EXtreme Gradient Boosting (XGBoost) Model Development

XGBoost

EXtreme Gradient Boosting (XGBoost) algorithm is an optimized distributed implementation of gradient boosted decision trees, designed for computational speed and higher performance. Since its initial release in 2014 (Chen and Guestrin, 2016), in the past few years, XGBoost has been applied to a number of biomedical problems.

As an implement machine learning algorithm under the gradient boosting framework, the starting point of XGBoost is decision trees. However, here, each tree is fitted to the residuals (prediction errors) of the previous tree in order to gradually minimize the deviations between the model and the observed target data. This is done by giving more weight to the poorly modeled cases. In contrast to the Random Forest model, the trees are thus not independent of each other. Besides the different random samples, this is additionally achieved by the fact that not all predictors are available for selection at each branching, but only a randomly chosen subset, and get exceptionally high performance for regression as well as classification tasks. Classification trees are used to identify the class/category within which the input variables would most likely fall, while regression algorithms are suitable for continuous variables, and the tree is used to predict the value.

XGBoost algorithm has gradient boosting at its core. However, unlike simple gradient boosting algorithms, the XGboost model takes a parallelization approach in the process of sequential addition of the weak learners, whereby proper utilization of the CPU core of the machine is utilized, leading to greater speed and performance (Santhanam, 2016). Moreover, it is a distributed and scalable computing method that is available for large datasets.

Moreover, one benefit of the gradient boosting model is that for different loss functions, new algorithms are not required to be derived; it is enough that a suitable loss function be chosen and then incorporated with the gradient boosting framework.

Model Training

We used XGBoost to train both classification and regression models, respectively; several predict models were built depending on data type.

For *k*-mers data, the occurrence times of each *k*-mer in each sample were counted, and we used all possible segments as features and mapped the number of *k*-mers to [0, 1] with Min–Max normalization. For SNPs data, features were characterized by binary number as zeros and ones of all mutation sites. The data were divided into training and test set as 8:2.

Our XGBoost models were set as tree-based structure (with ***booster*** = ***"gbtree"***), and GridsearchCV was applied for hyperparameter tuning. In order to prevent the XGBoost training process from generating too many trees, which causes the machine learning model to eventually overfit, we use fivefold cross-validation to select the most appropriate number of iterations; the value of booster_round is used as the num of XGBoost booster_round parameter, which is brought into the model training. Also, considering that our dataset is on the small

[1] https://www.ncbi.nlm.nih.gov/assembly/GCF_000240185.1

side, using cross-validation also allows training with as much data as possible.

We first trained the XGboost multi-classification model, with the objective parameter ***Multi: Softmax***. Input samples are fed into the generated XGBoost tree, and the leaf to which the sample belongs is found in each tree; the belonging weight is then added to obtain the predictions. As it is a multiclass classification model, we set 17 categories as classification labels to train the model, with a minimum MIC value of 0 and a maximum MIC value of 16, equally divided into 17 intervals. The prediction results are obtained by the ***softmax*** function, as probabilities of belonging to a certain MIC interval. For the regression model, the objective parameter of XGboost is ***Reg: Gamma***, as MIC values can be regarded as gamma-distributed. The MICs of each sample were used as label of model training.

To prevent the XGBoost training process from generating too many trees and causing the machine learning model to be overfitted, we use fivefold cross-validation to find the most appropriate number of iterations (***num_booster_round = "2000"***) to the model training. In addition, using cross-validation also allows us to use as much data as possible for training, considering our small dataset. Also, the maximum depth of the tree, ***max_depth***, was set to 6, and the proportion of random sampling, ***subsample***, is 0.6.

The accuracy of the model was determined by the absolute value of the difference between the log2-transform of the predicted values and the true values.

DNN Model Development

DNN

Deep learning is a concept for an approach to artificial intelligence called neural networks, and the DNN model is a basic deep learning framework. As a particular class of artificial neural networks with fully connected architecture, between the input and the output layer, there is an arbitrary number of hidden layers (Zador, 2019).

In principle, neural networks usually consist of four components: The input layer, the hidden layer(s), the output layer, and edges that connect the individual layers. More precisely, the edges connect individual nodes within the layers, whereby each transfer functions as a kind of container for a numerical value. The edges between the nodes have weights that define how the input is calculated across the edge to the next node. The arrangement of these components depends on the type and purpose of the network. Thus, the main difference between DNN and classical machine learning methods is the ability to process unstructured data through artificial neural networks (Dargan et al., 2019).

Model Training

To further improve the performance of MIC prediction, we assessed the importance of *k*-mers and SNPs, respectively, based on the previous XGBoost model. We ranked all *k*-mers and SNP features using f-score as standard, and we found that the f-score values of *k*-mers and SNP features that were ranked in top 40 were greater than 1, while the others were not that significant. Thus, for the DNN method, the top 40 most important *k*-mers and SNPs were selected as features for the deep learning-based modeling. We established the following three models to predict MIC value: *k*-mers model, SNPs model, and *k*-mers & SNPs model. Our overall work flow of MIC prediction modeling is shown in **Figure 1**.

The DNN model with *k*-mers and SNP inputs uses a ***Dense*** neural network framework, where the top 40 most important features for predicting MIC values are fed into a 128-unit Dense layer with a ***relu*** activation function to train the DNN model. Similarly, on the test set, the absolute value of the difference between the log2 transform of the predicted value and the true value is used as the basis for assessing the accuracy of the model.

In particular, for the *k*-mers & SNPs input, we use a combined Dense + LSTM model frame. More specifically, for the top 40 characteristic *k*-mers data selected by the previous model, input the Dense layer and then input the selected top 40 feature data from the SNP site into the LSTM layer. The Dense layer and the LSTM layer are combined as the model input to train the DNN model.

RESULTS

We first used the XGBoost classification model and made five predictions using KMER (110 samples * 559,494 *k*-mers features) and SNPs (110 samples * 164,138 SNPs features) data. For each experiment, we set different random states from 1 to 5. Similarly, the XGBoost regression model was used to make five times predictions for both *k*-mers and SNPs data. The random states parameter was taken from 1 to 5 in order to maintain consistency

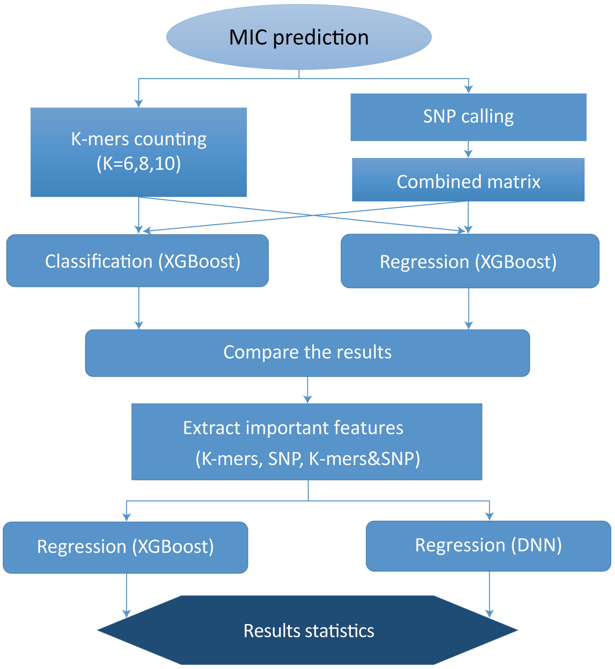


FIGURE 1 | Work flow of MIC prediction modeling.

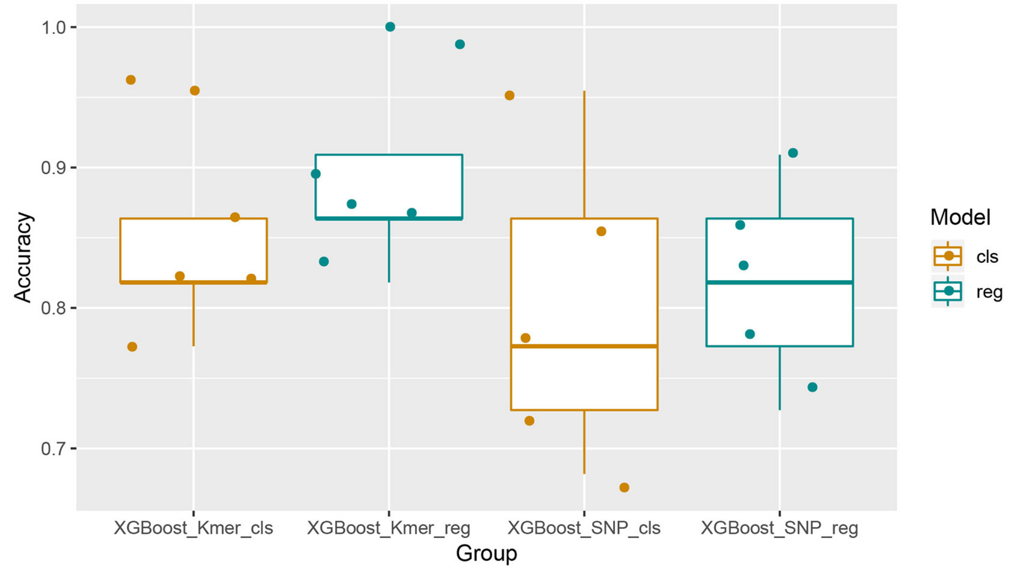


FIGURE 2 | Boxplots with jittered data points of XGBoost prediction accuracies for all features. It can be seen that the results of XGBoost regression are better than the classification and that XGBoost performed better with the *k*-mers characteristics than it with SNPs.

in the splitting of the dataset for comparative analysis of the results. A comparison of the prediction accuracies of the models was then performed. The Boxplot grouping in **Figure 2** shows the accuracy values for each of the five predictions, and **Table 1** shows their mean accuracy. From these results, it is clear that the XGBoost regression model predicts better than the classification model, for both *k*-mers and SNPs data. In addition, in terms of the input feature type, XGBoost predicted *k*-mers data with better accuracy than SNPs, possibly related to the fact that SNPs is a binary input of 0 and 1. The mean predictive accuracy of the XGBoost classification model for SNPs was 0.8, while the mean accuracy of the XGBoost regression model for *k*-mers reached 0.8909091.

The top 10 important features of the classification and regression models with *k*-mers and SNPs data were statistically analyzed, respectively, and presented in the bar chart in **Figure 3**. As can be seen from the figure, the top 10 features of the five attempts did not completely coincide, but some common features can be found. For example, for *k*-mers' classification model, CGACAGTCTC appears in all five runs, GACTCCTAGC appears four times in *k*-mers' regression model, and A2872728 and G17357 also appear four times each in SNPs' regression model.

To further optimize the model, the *k*-mers and SNPs top 40 feature datasets were taken for modeling and prediction by XGBoost regression and DNN regression, respectively. In order to enhance the reliability of the results, we used the XGBoost regression algorithm to model and predict all the features of *k*-mers and SNPs for another five times (the random_state parameter of the train_test_split function was taken from 6 to 10), and we also took their top 40 feature datasets for the XGBoost regression and DNN regression modeling. The top 40 feature datasets were also taken for the XGBoost regression and DNN regression modeling predictions.

TABLE 1 | Mean prediction accuracies of the XGBoost algorithm using all features of *k*-mers or SNPs (five times).

XGBoost	*k*-mers	SNPs
Classification	0.845	0.800
Regression	0.891	0.818

Next, we ran the XGBoost regression model 10 times, and for the top 40 feature dataset for each experiment, we ran the XGBoost regression prediction 10 times (random states from 1 to 10). The Boxplot grouping in **Figure 4** shows the accuracy values for each of the 100 predictions, and **Table 2** tallies their mean values. The XGBoost regressions for *k*-mers, SNPs, and *k*-mers & SNPs data had prediction accuracies of 0.9113636, 0.8522727, and 0.9127273, with the lowest predictive accuracy for SNPs and the best for *k*-mers & SNPs. Overall, the XGBoost regression model predicted the top 40 feature dataset better than the predictions for all feature datasets, for both *k*-mers and SNPs (**Tables 1**, **2**). We show the *y*-test and *y* predicted values for all 100 predictions and see that the predicted values largely fluctuate around the true values (**Figure 5**).

Similarly, for the DNN model, the top 10 important features selected by XGBoost were trained for a total of 100 times of random resolution, respectively. The Boxplot grouping in **Figure 6** shows the accuracy values of 1,000 times of prediction, and their average values are calculated in **Table 3**, and the test and predicted values for all 1,000 predictions are shown in **Figure 7**. Regressions for *k*-mers, SNPs, and *k*-mers & SNPs had prediction accuracies of 0.9189091, 0.8705455, and 0.9177273, respectively, with the lowest prediction accuracy for SNPs and very similar prediction accuracies for *k*-mers and *k*-mers & SNPs, all of which were relatively high.

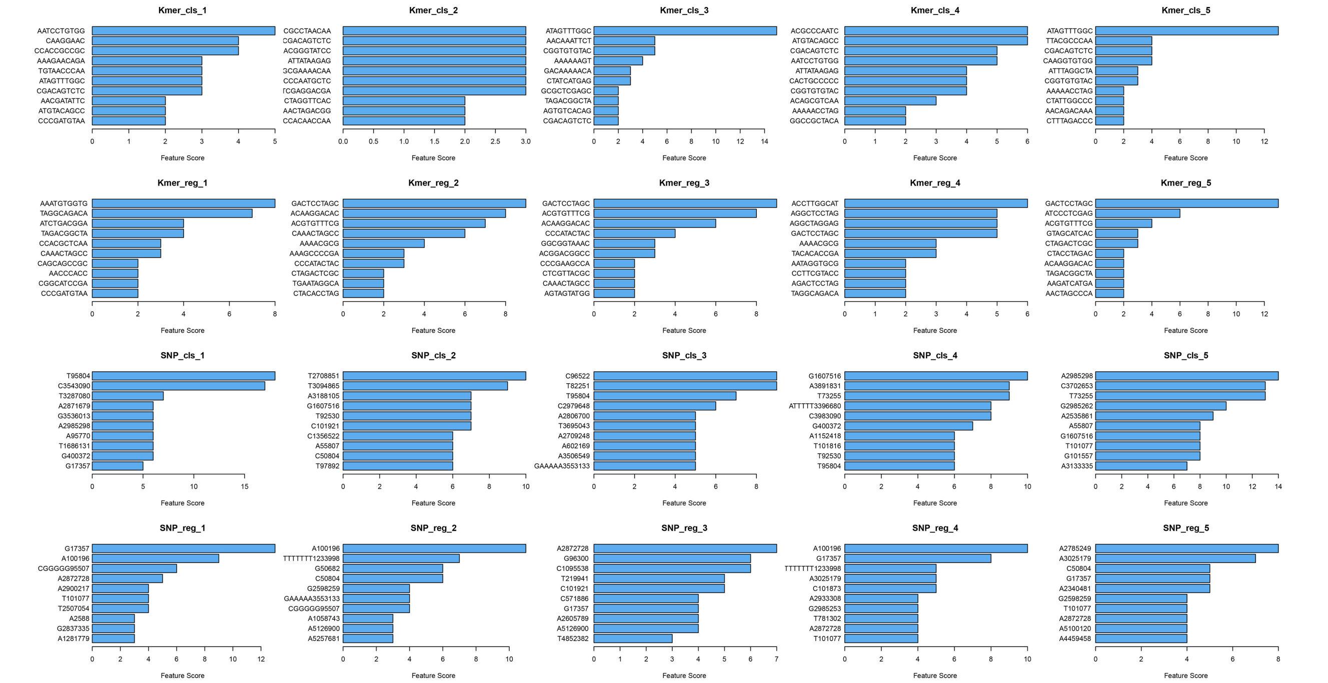


FIGURE 3 | Horizontal barplots for the top 10 important features of XGBoost classification and regression models.

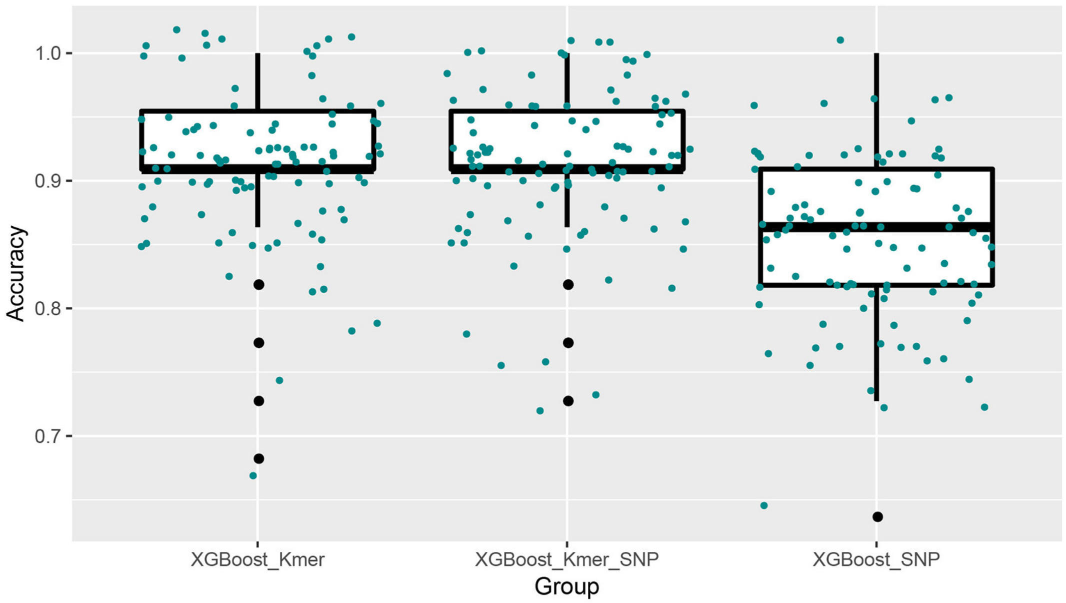


FIGURE 4 | Boxplots with jittered data points of XGBoost prediction accuracies for top 40 features. Since 10 × 10 = 100 modeling predictions were made, the results of 100 predictions in each box could be seen in XGBoost's comparison with *k*-mers, SNPs, and *k*-mers & SNPs.

For regression models, the mean square root of the error between the predicted and true values (RMSE) is usually used as a model evaluation metric, and the coefficient of determination (R^2) is used to indicate how well the model predicts the true value compared to the mean value model. We calculated the RMSE and R^2 values of our XGBoost and DNN models. For our XGBoost models, the RMSE values were 1.734, 2.781, and 1.717, and R^2 values were 0.860, 0.640, and 0.863, respectively (**Figure 5**). The RMSEs of the DNN models were 1.955, 2.179, and 2.045 and the R^2 values of the DNN model were 0.836, 0.796, and 0.820 (**Figure 7**). R^2 is an indicator used in regression models to evaluate the degree of agreement between the predicted value

TABLE 2 | Mean prediction accuracies of the XGBoost algorithm using top 40 features of *k*-mers or/and SNPs (10 × 10 times).

XGBoost (Top 40)	*k*-mers	SNPs	*k*-mers & SNPs
Regression	0.911	0.852	0.913

and the actual value, with a maximum value of 1. It can be seen that, overall, our models fit well.

In summary, our analysis showed that the XGBoost classification model reached over 80% prediction accuracy, and the model with *k*-mers data gave better results than SNPs inputs. Compared with the XGBoost classification model, the overall performance of the XGBoost regression model is improved (89.1 and 81.8% for *k*-mers and SNPs data, respectively). The MIC value is continuously distributed, and the effect of the regression model may be more realistic. DNN neural network

TABLE 3 | Mean prediction accuracies of the DNN algorithm using top 40 features of *k*-mers or/and SNPs (100 × 10 times training).

DNN (Top 40)	*k*-mers	SNPs	*k*-mers & SNPs
Regression	0.919	0.871	0.918

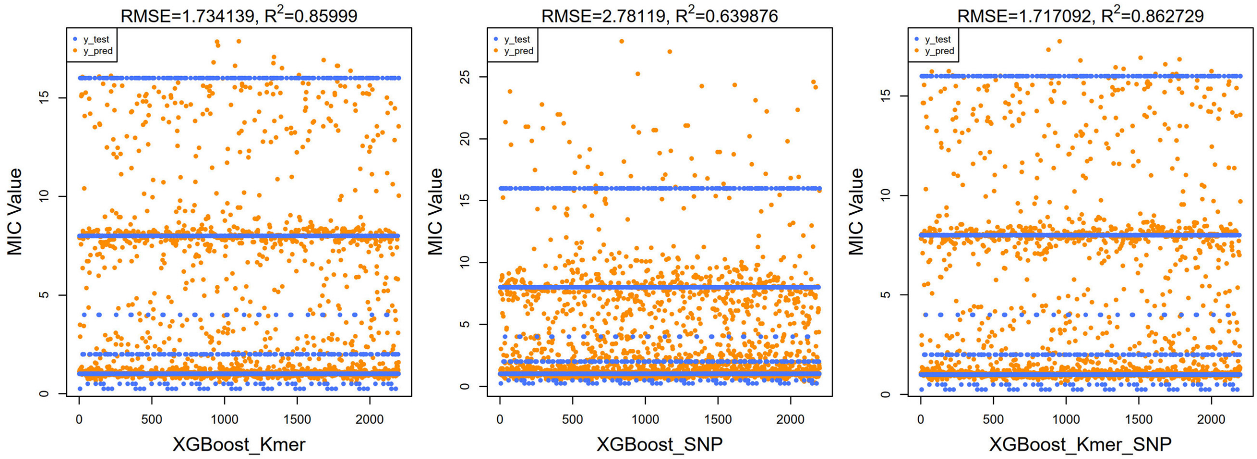


FIGURE 5 | Scatter plots of true test values and predicted values of MIC using XGBoost algorithm for top 40 features (left for *k*-mers, middle for SNPs, and right for *k*-mers & SNPs). As the original *y*-value was discrete, several horizontal lines were presented in the figure. The predicted values were clustered around these lines'; RMSE and R^2 values were also calculated and shown at the top of the figure.

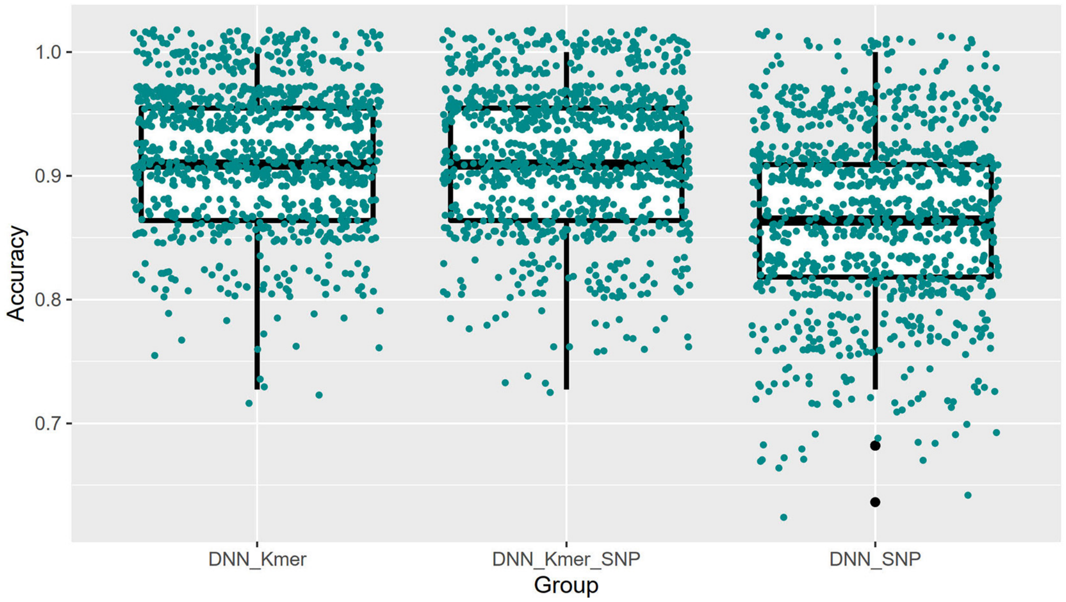


FIGURE 6 | Boxplots with jittered data points of DNN prediction accuracies for top 40 features (100 × 10 times training).

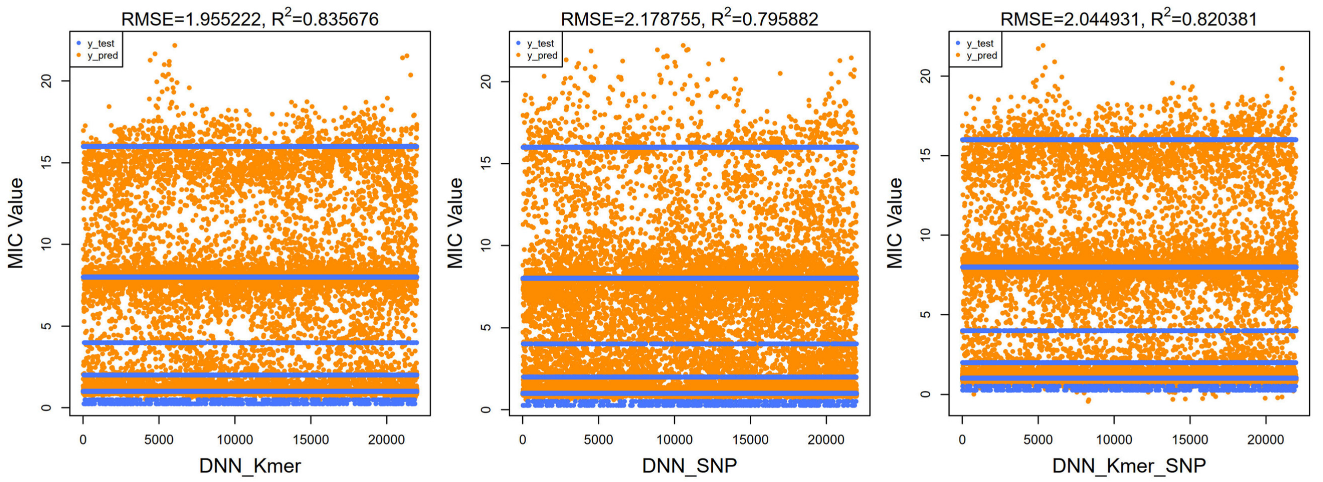


FIGURE 7 | Scatter plots of true test *y*-values and predicted *y*-values using DNN algorithm for top 40 features (100 × 10 times training). The predicted values were clustered around these lines; RMSE and R^2 values were also calculated and shown at the top of the figure.

models perform better in predicting MIC values with improved overall accuracy compared to XGBoost models. On the other hand, the *k*-mers and SNPs top 40 feature dataset was sufficient to obtain good prediction results (above 85% accuracy), with *k*-mers and mixed *k*-mers & SNPs features performing well and the DNN regression model performing better than the XGBoost regression approach.

DISCUSSION

Based on metagenomic data, in this study, sequence analysis was used to obtain SNPs information and nucleotide *k*-mers count information queue data; machine learning and deep learning methods were then applied to establish a prediction model for the MIC value of *K. pneumoniae*. By feature selection, we proposed a top 40 feature-based regression model, which had the best predictive performance of 91%.

First, according to Naha et al. (2021) and Okanda et al. (2020), we found that gene mutations may affect drug resistance of *Klebsiella*; thus, we tried to find the relevant sites affecting resistance by calling SNPs. After pre-processing the raw data by using biogenetics tools BWA, BCFTools, and SamTools, we obtained a matrix of mutation site and sample list. We took the mutated gene site as the features and built the machine learning model of classification and regression, respectively. We used 110 samples for prediction, and the prediction results above show that the mean accuracy of the SNPs classification model was 80% and the mean accuracy of the SNPs regression model was 81.81%, which shows that the performance of the regression model is better than the multi-classification model. Then, based on the method previously described by Nguyen et al. (2019), we created both XGBoost classification and regression models using *k*-mers counts as input features, respectively, and made MIC predictions for 110 samples. As described above, after five runs, we obtained a mean accuracy of 84.54% for the *k*-mers classification model and 89.09% accuracy for the *k*-mers regression model. This result again shows that the multi-classification model does not perform as well as the regression model. In addition, the prediction of MIC values using SNPs loci was less effective than that of *k*-mers prediction, which may be due to the fact that the input to the SNPs is binary data with only mutated (labeled as 1) and unmutated (labeled as 0) features, while the input to the *k*-mers counting model are continuous variables, making it more effective for regression model training.

To evaluate our model, we compared MIC prediction models built by related studies. In the study by ValizadehAslani et al. (2020), the authors used the XGBoost model with *k*-mers features, and the result shows an accuracy of around 91% in predicting the MIC value of meropenem against *K. pneumoniae*, which was close to our results. Another study by Nguyen et al. (2019) also used the XGBoost model to predict MICs for non-typhoidal *Salmonella*, resulting in an average accuracy of 90% without a large number of samples. We decided to try more advanced deep learning approach for prediction. As the K-mers and SNPs had too many feature values, and the neural network could not accept features with too high dimensions, we selected some of important features as the training data to avoid overfitting.

The XGBoost regression model gives a score of importance for each feature during the training process. We selected the top 40 highest scores from the *k*-mers and the SNPs regression model, respectively, and then we used these total 80 important features as a new dataset, to predict MIC values using both XGBoost and DNN algorithms. In consideration of training time and server capacity, we only use regression models for prediction.

Comparing the results in **Tables 2**, **3**, the DNN model performs better than the classical XGboost machine learning approach in predicting MIC values, with a slight improvement in both accuracy rates. However, the reason for the small improvement may be due to the fact that only important features were selected for training and the overall amount of sample size was relatively small. In addition, the prediction accuracy of the model improved by combining the significant features of *k*-mers and SNPs to produce a new dataset than training with a single type of feature.

We found the annotated.gff file of the reference genome from NCBI and the paper on the whole gene analysis of the reference genome HS11286 by the team of Liu (Liu et al., 2012); the *K. pneumoniae* resistance genes were found from this paper and we identified loci belonging to these gene fragments from important features in the SNPs model. The pKPHS3 was mentioned in the study (Liu et al., 2012) as possessing 13 important resistance determinants, such as tetG, cat, sul1, dfra12, aac(3)-Ia, and aph. Genes were found among the important features of our SNPs, such as site T37808, which belongs to the tetG gene family, an important gene family that influences tetracycline resistance. This demonstrates that the important feature values obtained from our model training may help us to understand the reasons for the development of resistance, and why there are anti-tetracycline resistance genes present due to the presence of tra isoconjugate transfer genes in pKPHS2 and pKPHS3, which is the type of gene that causes resistance to spread between genera (Liu et al., 2012). Moreover, meropenem belongs to the class of beta-lactam antibiotics, which are classified as carbapenems. According to Reyes et al. (2019), the most common resistance mechanism of *K. pneumoniae* to carbapenem antibiotics is the production of enzymes with carbapenemase activity, which hydrolyze beta-lactam antibiotics, while we also identified mutations in the beta-lactamase gene from important features in SNPs models, such as C1114518 and G1114674; i.e., mutations in the beta-lactamase gene may be responsible for the high MIC values.

In summary, we found that there are still a lot of genes in *Klebsiella* that belong to hypothetical proteins, and the loci we derived from this study can help to annotate and study these hypothetical proteins. Furthermore, in clinical practice, deep learning-based modeling and prediction by selecting important feature values can significantly improve detection efficiency compared to experimental methods of measuring MIC values, providing doctors with a faster access to information on patient resistance for drug administration and improving the effectiveness of antibiotic use, enabling patients to receive medication promptly. It also reduces the cost of the experiment.

ADDITIONAL INFORMATION

CentOS Linux release 7.2.1511 (Core)

Linux version 3.10.0-327.el7.x86_64 (builder@kbuilder.dev.centos.org) (gcc version 4.8.3 20140911 (Red Hat 4.8.3-9) (GCC)

jupyter lab version 0.34.9

Python 3.7.2

DATA AVAILABILITY STATEMENT

The metagenomic sequence data included in this study can be found in the NCBI SRA (BioProject accession numbers PRJNA376414, PRJNA386693, and PRJNA396774).

AUTHOR CONTRIBUTIONS

JG and JL conceived ideas and designed the study. AY, JG, XJW, and XLW wrote the manuscript. ZML and RJ performed the bioinformatics analysis. RT and ZQL constructed the machine learning models. All authors read or revised the manuscript and approved the final version.

FUNDING

This study was mainly supported by the National Key R&D Program of China (2018YFE0102400) and secondly by the Talent Development Project for Three-year Action Plan of Shanghai Public Health System Construction (GWV-10.2-XD03 and GWV-10.2-YQ50) and the project Clinical Treatment Optimization Decision System Based on Deep Reinforcement Learning by Shanghai Municipal Commission of Economy and Informatization (2020-RGZN-02039). XJW received funding from the Scientific Instrument Application Methods Project of Shanghai Science and Technology Innovation Action Plan (No. 19142200800).

ACKNOWLEDGMENTS

We would like to thank Yingxia Pan and Xiaoming Li who contributed to data pre-processing and programming.

REFERENCES

Baldwin, C. M., Lyseng-Williamson, K. A., and Keam, S. J. (2008). Meropenem: a review of its use in the treatment of serious bacterial infections. *Drugs* 68, 803–838. doi: 10.2165/00003495-200868060-00006

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

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., et al. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* 10:giab007. doi: 10.1093/gigascience/giab007

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, (San Francisco CA: Association for Computing Machinery).

Dargan, S., Kumar, M., Ayyagari, M. R., and Kumar, G. (2019). A survey of deep learning and its applications: a new paradigm to machine learning. *Arch. Comput. Methods Eng.* 27, 1071–1092.

Effah, C. Y., Sun, T., Liu, S., and Wu, Y. (2020). *Klebsiella pneumoniae*: an increasing threat to public health. *Ann. Clin. Microbiol. Antimicrob.* 19:1. doi: 10.1186/s12941-019-0343-8

Eyre, D. W., De Silva, D., Cole, K., Peters, J., Cole, M. J., Grad, Y. H., et al. (2017). WGS to predict antibiotic MICs for *Neisseria* *gonorrhoeae*. *J. Antimicrob. Chemother.* 72, 1937–1947. doi: 10.1093/jac/dkx067

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

Li, Y., Metcalf, B. J., Chochua, S., Li, Z., Gertz, R. E. Jr., Walker, H., et al. (2016). Penicillin-binding protein transpeptidase signatures for tracking and predicting beta-lactam resistance levels in *Streptococcus pneumoniae*. *mBio* 7:e00756-16. doi: 10.1128/mBio.00756-16

Li, Y., Metcalf, B. J., Chochua, S., Li, Z., Gertz, R. E. Jr., Walker, H., et al. (2017). Validation of beta-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. *BMC Genomics* 18:621. doi: 10.1186/s12864-017-4017-7

Liu, P., Li, P., Jiang, X., Bi, D., Xie, Y., Tai, C., et al. (2012). Complete genome sequence of *Klebsiella pneumoniae* subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *J. Bacteriol.* 194, 1841–1842. doi: 10.1128/JB.00043-12

Long, S. W., Olsen, R. J., Eagar, T. N., Beres, S. B., Zhao, P., Davis, J. J., et al. (2017). Population genomic analysis of 1,777 extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates, Houston, Texas: unexpected abundance of clonal group 307. *mBio* 8:e00489-17. doi: 10.1128/mBio.00489-17

Naha, S., Sands, K., Mukherjee, S., Saha, B., Dutta, S., and Basu, S. (2021). OXA-181-like carbapenemases in *Klebsiella pneumoniae* ST14, ST15, ST23, ST48, and ST231 from septicemic neonates: coexistence with NDM-5, resistome, transmissibility, and genome diversity. *mSphere* 6:e01156-20. doi: 10.1128/mSphere.01156-20

Navon-Venezia, S., Kondratyeva, K., and Carattoli, A. (2017). *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol. Rev.* 41, 252–275. doi: 10.1093/femsre/fux013

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., et al. (2018). Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.* 8:421. doi: 10.1038/s41598-017-18972-w

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., and Stevens, R. L. (2019). Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* 57:e01260-18.

Okanda, T., Haque, A., Koshikawa, T., Islam, A., Huda, Q., Takemura, H., et al. (2020). Characteristics of carbapenemase-producing *Klebsiella pneumoniae* isolated in the intensive care unit of the largest tertiary hospital in Bangladesh. *Front. Microbiol.* 11:612020. doi: 10.3389/fmicb.2020.612020

Pataki, B. A., Matamoros, S., van der Putten, B. C. L., Remondini, D., Giampieri, E., Aytan-Aktug, D., et al. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning. *Sci. Rep.* 10:15026. doi: 10.1038/s41598-020-71693-5

Pesesky, M. W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C. D., et al. (2016). Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative *Bacilli* from whole genome sequence data. *Front. Microbiol.* 7:1887. doi: 10.3389/fmicb.2016.01887

Reyes, J., Aguilar, A. C., and Caicedo, A. (2019). Carbapenem-Resistant *Klebsiella pneumoniae*: microbiology key points for clinical practice. *Int. J. Gen. Med.* 28, 437–446. doi: 10.2147/IJGM.S214305

Santhanam, R. (2016). Comparative study of XGBoost4j and gradient boosting for linear regression. *Int. J. Control Theory Appl.* 9, 1131–1142.

Spagnolo, A. M., Orlando, P., Panatto, D., Perdelli, F., and Cristina, M. L. (2014). An overview of carbapenem-resistant *Klebsiella pneumoniae*: epidemiology and control measures. *Rev. Med. Microbiol.* 25, 7–14. doi: 10.1097/MRM.0b013e328365c51e

ValizadehAslani, T., Zhao, Z., Sokhansanj, B. A., and Rosen, G. L. (2020). Amino acid k-mer feature extraction for Quantitative Antimicrobial Resistance (AMR) prediction by machine learning and model interpretation for biological insights. *Biology* 9:365. doi: 10.3390/biology9110365

Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., et al. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* 34, 1666–1671. doi: 10.1093/bioinformatics/btx801

Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W., et al. (2019). DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. *Bioinformatics* 35, 3240–3249. doi: 10.1093/bioinformatics/btz067

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* 10:3770.

Conflict of Interest: JG, AY, RT, XJW, and RJ are employed by Shanghai Biotecan Pharmaceuticals Co., Ltd. and Shanghai Zhangjiang Institute of Medical Innovation.

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.

TECHNOLOGY AND CODE
published: 06 October 2021
doi: 10.3389/fmicb.2021.714836

Graph-Based Approaches Significantly Improve the Recovery of Antibiotic Resistance Genes From Complex Metagenomic Datasets

Daria Shafranskaya[1,2], *Alexander Chori*[2,3] and *Anton Korobeynikov*[1,2]*

[1] Scientific Center for Information Technologies and Artificial Intelligence, Sirius University of Science and Technology, Sochi, Russia, [2] Center for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, Russia, [3] ITMO University, Saint Petersburg, Russia

The lack of control over the usage of antibiotics leads to propagation of the microbial strains that are resistant to many antimicrobial substances. This situation is an emerging threat to public health and therefore the development of approaches to infer the presence of resistant strains is a topic of high importance. The resistome construction of an isolate microbial species could be considered a solved task with many state-of-the-art tools available. However, when it comes to the analysis of the resistome of a microbial community (metagenome), then there exist many challenges that influence the accuracy and precision of the predictions. For example, the prediction sensitivity of the existing tools suffer from the fragmented metagenomic assemblies due to interspecies repeats: usually it is impossible to recover conservative parts of antibiotic resistance genes that belong to different species that occur due to e.g., horizontal gene transfer or residing on a plasmid. The recent advances in development of new graph-based methods open a way to recover gene sequences of interest directly from the assembly graph without relying on cumbersome and incomplete metagenomic assembly. We present GraphAMR—a novel computational pipeline for recovery and identification of antibiotic resistance genes from fragmented metagenomic assemblies. The pipeline involves the alignment of profile hidden Markov models of target genes directly to the assembly graph of a metagenome with further dereplication and annotation of the results using state-of-the art tools. We show significant improvement of the quality of the results obtained (both in terms of accuracy and completeness) as compared to the analysis of an output of ordinary metagenomic assembly as well as different read mapping approaches. The pipeline is freely available from https://github.com/ablab/graphamr.

Keywords: antibiotic resistance, assembly graphs, metagenome, profile hidden Markov model, computational pipeline

Edited by:
Liang Wang,
Institut Pasteur of Shanghai (CAS), China

Reviewed by:
Yanni Sun,
City University of Hong Kong, Hong Kong, SAR China
Rosa Del Campo,
Ramón y Cajal Institute for Health Research, Spain
Lateef Salam,
Elizade University, Nigeria

***Correspondence:**
Anton Korobeynikov
a.korobeynikov@spbu.ru

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

Received: *25 May 2021*
Accepted: *09 September 2021*
Published: *06 October 2021*

Citation:
Shafranskaya D, Chori A and Korobeynikov A (2021) Graph-Based Approaches Significantly Improve the Recovery of Antibiotic Resistance Genes From Complex Metagenomic Datasets.
Front. Microbiol. 12:714836.
doi: 10.3389/fmicb.2021.714836

INTRODUCTION

Antimicrobial resistance (AMR) is a global health crisis resulting from widespread and uncontrolled use of antibiotics (Brown and Wright, 2016). Therefore, the use of genome sequencing as a surveillance tool for AMR molecular epidemiology is growing, and the development of new computational approaches is an important task (McArthur and Wright, 2015).

Certainly, there are many tools developed recently for AMR prediction and analysis from WGS data (Boolchandani et al., 2019). In general, all these tools could be splitted into two groups: ones that use raw sequencing reads as input, such as SRST2 (Inouye et al., 2014) that use paired-end-aware short read aligner to align reads to reference databases or first splitting reads into k-mers and then aligning them to databases such as KmerResistance (Clausen et al., 2016). Another group of tools that use assembled genome fragments includes Abricate (https://github.com/tseemann/abricate), RGI (Jia et al., 2017), Resfinder (Bortolaia et al., 2020) among the others. ARIBA (Hunt et al., 2017) and RGI (Jia et al., 2017) could utilize both reads and assembled fragments, however, this does not change in general their approach for AMR prediction.

The natural limitation of any read-based approach is the input read length and therefore the precision of such approach might suffer from the truncated read-gene mappings (depending on the target AMR gene length). **Figure 1** shows the distribution of AMR gene lengths in the NCBI AMR database (Feldgarden et al., 2019) with the majority of genes, namely 93%, that are more than 300 base pairs long. Given that typically the reads produced by short reads technologies are within 100–300 bp length, the read-based methods would need to cope with incomplete alignments of reads to AMR databases or additional techniques (e.g., overlapping paired-end reads) would be required in order to correctly cover the genes of interest.

Another approach involves the use of sequences obtained from raw reads after the genome assembly process. Genome assembly may overcome the difficulties connected with the lengths of short reads and allows for reconstruction of fuller gene sequences, however it still has some limitations on its own. Possible issues include possible assembly artifacts, increased computational processing time, etc. Nonetheless, all these issues could certainly be detected, most of them solved in automatic fashion and therefore AMR prediction on top of microbial isolate assembly could be considered a mostly solved problem.

However, the overall situation is much worse when one would need to analyse a resistome from an environmental sample, such as water metagenome, or human-associated sample, e.g., gut metagenome. Such assemblies are often very fragmented due to vastly different species abundance, presence of multiple strains, interspecies repeats that arise from conservative genes or genes that underwent horizontal transfer, etc. (Lapidus and Korobeynikov, 2021). Even more, metagenomic assemblers typically yield a consensus assembly (Nurk et al., 2017) with collapsed strain variations complicating the necessary prediction.

As a result, AMR prediction from metagenomic assembly can show quite low specificity with many important AMR genes unnoticed (Maguire et al., 2020).

To support this claim we analyzed wastewater and urban surface metagenomes in Singapore from Ng et al. (2017) that originally used a read-based approach to construct a resistome. First example deals with bla$_{\text{IMP}}$ beta-lactamase gene that according to Ng et al. (2017) was absent in the sample. This is not unexpected given the length of bla$_{\text{IMP}}$ gene cassette of 741 bp (encoding 246 amino acid polypeptide) (Silva et al., 2002) that certainly could escape from read-based analysis. Furthermore, additional analysis shows that the complete sequence of bla$_{\text{IMP}}$ is absent in assembled scaffolds as well, however the bla$_{\text{IMP}}$ gene sequence is definitely present in the sample. This phenomenon could be easily explained by examining the assembly graph. **Figure 2** shows that the gene sequence of bla$_{\text{IMP}}$ is contained in 10 edges of the assembly graph and 2 scaffolds, hindering assembly-based analysis.

Sometimes, the gene of interest could be found in contigs, however, when multiple variants are present, not all of them could be easily identified from the contigs alone. **Figure 3** shows different variants of the bla$_{\text{CTX}-\text{M}}$ gene in the assembly graph of the same sample from Ng et al. (2017). We note that CTX-M-15 variant of the gene is residing on the single contig and therefore could be easily identified. However, CTX-M-9 and CTX-M-14 variants differ only by 2 amino acids and therefore assembler

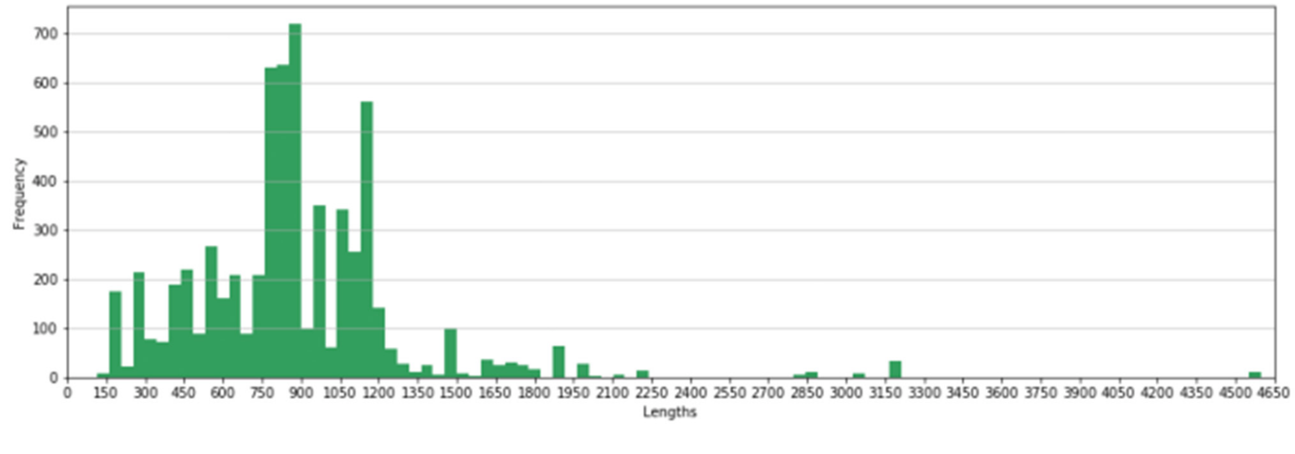


FIGURE 1 | Distribution of AMR gene lengths in the NCBI AMR database.

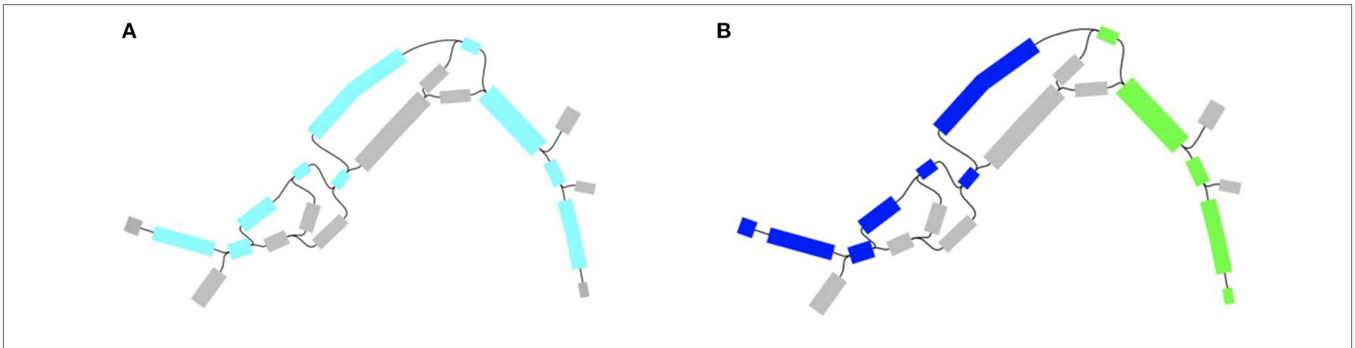


FIGURE 2 | bla_{IMP} sequence in the assembly graph. **(A)** Gene sequence scattered over 10 edges, **(B)** Gene sequence is splitted across two scaffolds.

is unable to separate them: CTX-M-14 is scattered across 3 contigs that are joined into single scaffold with gaps and CTX-M-9 is completely unassembled as its variation with respect to CTX-M-14 is reported as separate short contigs.

The examples shown above suggest the use of the assembly graph for AMR prediction from complex metagenome sequences since it is the assembly graph rather than set of contigs that represents the "complete" metagenomic assembly result. Even more, metagenomic assemblers provide both so-called strain assembly graph with strain variants preserved and consensus assembly graph with strain variants collapsed (Lapidus and Korobeynikov, 2021), so one could control the tradeoff between specificity and complexity of the task.

Finally, to show the possible performance gains from assembly graph-based approaches we used PathRacer (Shlemov and Korobeynikov, 2019), a tool that performs profile HMM alignment to assembly graphs, to align NCBI-AMR (Feldgarden et al., 2019) set of AMR profile HMMs to the assembly graphs of samples from Ng et al. (2017) and counted the fraction of HMM hits that are not residing on the single scaffold. **Figure 4** shows the results obtained. Overall, more than 30% of all HMM hits are not contained in the single scaffold supporting the idea of using graph-based tools for AMR prediction.

Motivated by the data shown above we are presenting GraphAMR—a novel computational pipeline that utilizes assembly graph of a metagenome for AMR prediction. GraphAMR uses state-of-art tools to align profile HMMs representing AMR gene families, extract the sequences of graph edges that contain HMM hits and uses well-known AMR-prediction tools to further annotate the obtained sequences.

PIPELINE ARCHITECTURE

GraphAMR is a pipeline specifically designed for recovery and identification of antibiotic resistance genes from fragmented metagenomic assemblies. Briefly, it uses state-of-the-art assembly graph analysis methods to extract putative AMR gene sequences from the graph, dereplicates them and delegates the task of actual prediction to the well-known AMR analysis tools in the field.

The pipeline is implemented using the Nextflow framework (Di Tommaso et al., 2017; Ewels et al., 2020) that enables scalable, reproducible and efficient computational workflow. As a result, the pipeline supports e.g., job submissions on computational clusters and cloud systems, resume, and notification straight out of the box.

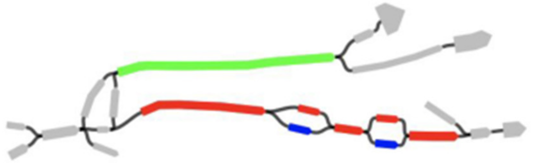


FIGURE 3 | bla_{CTX-M} paths and their neighborhood. Green path corresponds to CTX-M-15 variant; blue and red corresponds to CTX-M-9 and CTX-M-14, respectively.

The pipeline has four steps: (optional) metagenomic *de novo* assembly, alignment of AMR profile HMM to the resulting assembly graph, detection, and clustering of putative AMR ORFs and annotation of representative AMR sequences (**Figure 5**). The first step (assembly) can be skipped, should the assembly graph in the GFA (https://github.com/GFA-spec/GFA-spec) format be provided as an input. Such assembly graphs are readily produced by genome and metagenome assemblers including SPAdes (Prjibelski et al., 2020), metaSPAdes, and MEGAHIT (Li et al., 2015).

De novo Assembly

If reads are provided as input, the first step will be quality control and metagenomic assembly. Sequences QC is performed via FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The resulting HTML report shows summary graphs with main characteristics for quality assessment. Metagenome assembly is done via metaSPAdes (Nurk et al., 2017) and the resulting assembly graph is used for further analysis.

Profile HMM or AA Sequence Alignment to Assembly Graph

This is the key step of the pipeline as putative AMR gene sequences are extracted directly from the assembly graph. For this the pipeline utilizes Pathracer (Shlemov and Korobeynikov, 2019), a state-of-the-art tool for alignment of HMMs and

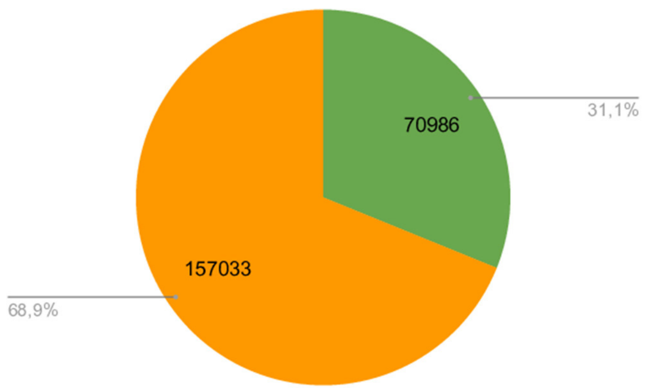


FIGURE 4 | Number of total HMM alignments to graph. The orange section shows the number of HMM hits residing on the scaffolds, and the green section shows the number of HMM hits possibly scattered over multiple scaffolds.

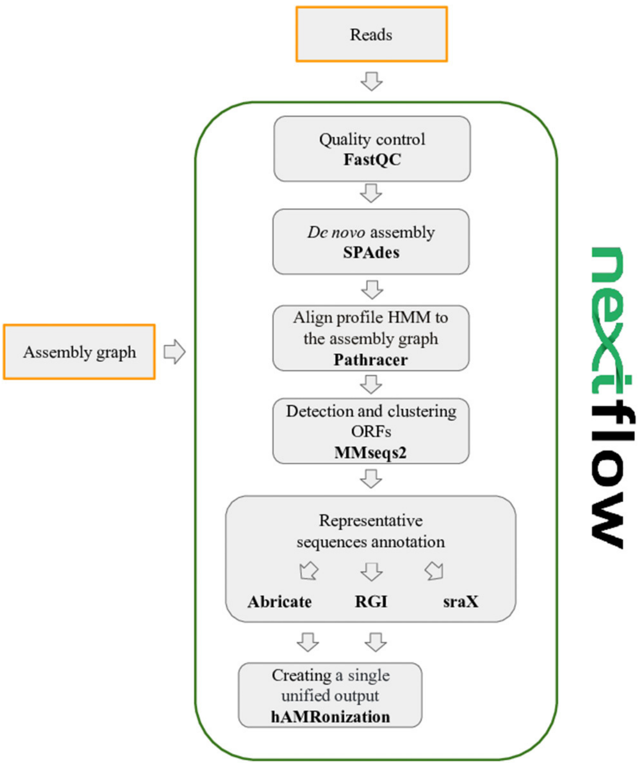

Reads
Quality control
FastQC
De novo assembly
SPAdes
Assembly graph
Align profile HMM to the assembly graph
Pathracer
Detection and clustering ORFs
MMseqs2
Representative sequences annotation
Abricate
RGI
sraX
Creating a single unified output
hAMRonization
nextflow

FIGURE 5 | General GraphAMR pipeline scheme.

AA sequences to assembly graph. By default, the NCBI AMR (Feldgarden et al., 2019) profile HMMs are used, but they could be replaced by the custom HMMs or gene AA sequences if necessary. Pathracer produces the set of most probable paths traversed by a HMM through the whole assembly graph (by default, up to top 100 by score non-redundant paths, e.g., those that are not proper suffixes or prefixes of each other, are reported). This effectively solves the problem of fragmented metagenome assemblies as all possible HMM paths (spanned over multiple contigs) are reported including possible variations due to multiple strains present, interspecies repeats, etc.

The major caveat here is that HMM alignment does not yield the complete gene sequence, since, for example, HMM could be built from the truncated seed alignment, or the alignment itself could be clipped on the ends. To solve this problem, instead of alignment itself, we extract the sequence of graph edges that contain the alignment of interest, effectively extending the alignment until the edge boundaries.

The output of this stage is the set of unique edge sequences of the assembly graph containing the alignments of profile HMMs of AMR genes.

In addition to HMMs, the pipeline also allows alignment of amino-acid sequences to the graph enabling the use of such AMR databases as CARD (Jia et al., 2017) or ResFinder (Bortolaia et al., 2020) directly. To enable the use of such databases, PathRacer internally builds a "proxy" HMM, so that the alignment of this HMM would be equivalent to the alignment of the original sequence using BLOSUM62 scoring matrix.

Dereplication

The output of the previous step might be redundant due to strain variations, but more because different edge sequences through the assembly graph might yield the same set of genes in the case when alignment ends in the node of the graph (recall that assembly graph is a de Bruijn graph, where subsequent edges overlap by a k-mer) or if there are multiple paths due to synonymous substitutions. To dereplicate the results, the complete ORFs are extracted and further clustered at 90% AA IDY using MMseqs2 (Steinegger and Söding, 2017). The output of this step is the set of representative sequences of the resulting clusters. The dereplication and clustering could be skipped via setting the IDY clustering threshold as 100%.

Annotation

There is no need to design a completely new AMR prediction approach given that the major challenges of obtaining putative AMR sequences from fragmented metagenome assemblies are solved via the proper utilization of the assembly graph. Therefore, this step delegates the task of final AMR prediction, annotation, and result generation to state of the art tools that are well-known and respected by the bioinformatics community. The pipeline passes the output of the dereplication stage to abricate (https://github.com/tseemann/abricate), sraX (Panunzi, 2020), and rgi (Jia et al., 2017). The results are further combined and summarized by hAMRonize tools (https://github.com/pha4ge/hAMRonization).

RESULTS

Usage

The pipeline is implemented in Nextflow and therefore requires Nextflow to be installed in order to be used. For the full reproducibility, the use of Nextflow-supported package manager such as Conda is advised. GraphAMR will automatically pull the necessary versions of the tools used in the pipeline when using

TABLE 1 | Abricate predicted AMR gene sequence counts in the URBAN dataset.

Sample ID	40	41	42	43	44	45	46	47	48	49	50	51	52
Contigs	92	93	2	0	4	9	57	0	91	3	11	78	66
HMM Paths	169	163	2	0	4	8	100	0	131	3	22	122	142
Clustered ORFs (90%)	103	98	2	0	4	8	60	0	96	3	11	91	80
Clustered ORFs (95%)	105	105	2	0	4	8	61	0	100	3	11	92	81
Clustered ORFs (100%)	135	126	2	0	4	8	75	0	112	3	14	107	116

Compared are assembled contigs, unclustered HMM paths, and clustered ORFs at different levels of IDY's. Columns are named by the last two digits of SRA accession number.

one of the supported container engines. The typical steps to run the pipeline for the first time are as follows:

1. Install nextflow (https://nf-co.re/usage/installation)
2. Install any Nextflow-supported container engines, such as conda (https://conda.io/miniconda.html)
3. Download the pipeline and test it on a minimal dataset with a single command: nextflow run ablab/graphamr -profile test, conda
4. Start running your own analysis:
   a. Typical command for analysis starting from reads (NCBI AMR database is used by default):
   ```
   nextflow run ablab/graphamr -profile conda --reads ‘*_R{1,2}.fastq.gz’
   ```
   b. Typical command for analysis starting from assembly graph (NCBI AMR database is used by default):
   ```
   nextflow run ablab/graphamr -profile conda --graph ‘assembly_graph_with_scaffolds.gfa’
   ```
   c. Typical command for analysis starting from assembly graph with one of pre-defined AMR databases:
   ```
   nextflow run ablab/graphamr -profile conda --graph ‘assembly_graph_with_scaffolds.gfa’ --db [‘ncbi_AMR_HMM’, ‘card_AA’]
   ```

More examples, description of other command line options and produced results are available from the "Usage/Results" section of documentation in GraphAMR github repository.

Example Results

To demonstrate the performance of graph-based approach for AMR discovery we benchmarked GraphAMR pipeline on two different environmental datasets using two different databases: NCBI AMR HMMs and amino acid sequences from CARD.

URBAN is a collection of urban wastewater datasets from Ng et al. (2017). Raw sequence reads were downloaded from the NCBI short read archive (SRA) under accession numbers SRR5997540–SRR5997552 and analyzed using the pipeline. For the sake of simplicity only AMR predictions by Abricate are shown. **Table 1** contains the predicted AMR gene counts predicted from metagenomic assembly scaffolds, unclustered HMM paths and HMM paths dereplicated, and clustered at different IDY's %. The results of the pipeline using amino acids are presented in **Table 2**.

TABLE 2 | Abricate predicted unique AMR gene sequence counts in the URBAN dataset using amino-acid sequences from CARD v3.1.2 or HMMs from NCBI AMR to align to a graph.

Sample ID	40	41	42	43	44	45	46	47	48	49	50	51	52
AA	96	89	2	0	4	8	59	0	89	3	11	84	74
HMM	94	89	2	0	4	8	59	0	89	3	10	82	74

Columns are named by the last two digits of SRA accession number.

The resulting AMR presence heatmap as produced by RGI is available as **Supplementary Figure 1**. The running time, physical memory usage and CPU usage and graph size information presented in the **Supplementary Figure 1** and **Table 1**, respectively.

We note that HMM paths represent unique path sequences over the assembly graph and might be redundant: two different paths in the graph may yield the same amino acid gene sequence, for example, due to synonymous mutations or if the alignment ends in the node of the graph since edges have overlapping k-mers. This explains the higher number of predicted AMR gene sequences obtained from bare HMM paths as compared to dereplicated or clustered ORFs.

The sample SRR5997545 looks like an outlier in **Table 1**, as the number of predicted AMR genes out of contigs is higher than from the assembly graph. The difference is caused by the short hit that resides on the isolated edge of the assembly graph. The hit itself covers only 73% of the HMM. By default Pathracer uses the strict threshold and does not report hits that are shorter than 90% of HMM length (we expect fuller HMM matches from the assembly graph as compared to contig sequences). To allow inclusion of such sequences should they be necessary we added a special flag to the pipeline that allows a user to choose the desired HMM coverage threshold.

To further compare the assembly graph-based approach with the read-based one we run SRST2 on the same collection of datasets. **Table 3** contains the predicted unique AMR gene counts from raw reads as detected by SRST2 and clustered HMM paths from GraphAMR. SRST2 uses a custom AMR database that was derived from CARD v3.0.8. To ensure fair comparison we run GraphAMR pipeline and Abricate using the database that was used by SRST2.

Table 3 clearly shows the advantage of the graph-based approach since more AMR gene sequences were predicted in

TABLE 3 | Predicted unique AMR gene sequence counts in raw reads of URBAN as detected by SRST2 vs. GraphAMR predictions from the assembly graph.

Sample ID	40	41	42	43	44	45	46	47	48	49	50	51	52
SRST2	59	55	6	0	2	6	36	0	59	2	8	54	44
GraphAMR	90	83	1	0	3	7	52	0	82	3	10	79	68

AMR annotation was done via Abricate. All tools used the CARD_v3.0.8_SRST2 database. Columns are named by the last two digits of SRA accession number.

almost all samples as compared to the read-based approach. Still, there is one notable outlier: in SRR5997542 sample SRST2 predicted 5 more AMR genes. Further detailed analysis revealed that these hits are likely spurious: the sequences themselves are fragmented on the assembled graph and the graph edges are isolated (see **Supplementary Figure 2**).

SOIL is groundwater metagenome sample SRR8931193 from Smith et al. (2019). Abricate predicted 12 AMR genes from clustered HMM paths and 13 from assembled scaffolds. Two gene sequences [vanR-O and ant(6)-Ib] genes were found only on scaffolds and tet(X) was detected by GraphAMR only. Assembly graph analysis revealed that ant(6)-Ib gene sequence is split into two parts located on two isolated edges. vanR-O hit covered only 30% of the corresponding sequence and is likely spurious.

DISCUSSION

As **Tables 1–3** and **Figure 4** show, the results of AMR gene prediction even on moderately-complex metagenomes could be significantly affected by fragmented assemblies. The use of assembly graph-based approaches is far superior in terms of recovery of fuller AMR gene sequences even from fragmented metagenomes. Not only could it result in more putative AMR sequences detected, but as comparison with read-based approaches shows, the results are more reliable. Graph-based approach allows to filter out the spurious alignments using both hit length (the fraction of the gene sequence length covered by a hit) and graph topology (short hits located on isolated edges are likely spurious) that results in AMR gene sequences that are both longer (hit could span multiple edges and interspecies repeats) and trustworthy (located on the edges of the graph that are connected to the rest of the assembly).

Another important task that could be solved using the assembly-graph based approach is AMR host association: sometimes it is not enough simply to detect the gene sequences, but also associate them with the particular species. This task is quite complex in case of metagenomic assemblies as a dedicated procedure called “binning” is required. However, typically binners ignore short contigs (shorter than 2–5 kbp) and therefore further detection of AMR gene sequences from MAGs could be quite limited (Maguire et al., 2020). Graph-based approach allows to circumvent this problem as one could trace the detected AMR sequences back to the edges of the assembly graph and then to the corresponding MAGs performing the required species identification. The challenge here certainly is dealing with interspecies repeats and/or plasmids or otherwise transferred genes, however, the assembly graph provides a solid foundation for such downstream analysis.

GraphAMR could be used to improve the present results of AMR prediction of a metagenomic assembly if the assembly graph output was preserved, otherwise the pipeline allows for seamless reassembly and AMR prediction starting from the input sequencing reads.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: NCBI SRA: SRR5997540–SRR5997552, SRR8931193.

AUTHOR CONTRIBUTIONS

AK contributed to conception and design of the study. AK, AC, and DS implemented the pipeline. AK and DS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This reported study was funded by RFBR, project number 19-34-51017.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.714836/full#supplementary-material

REFERENCES

Boolchandani, M., D’Souza, A. W., and Dantas, G. (2019). Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.* 20, 356–370. doi: 10.1038/s41576-019-0108-4

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., et al. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* 75, 3491–3500. doi: 10.1093/jac/dkaa345

Brown, E. D., and Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature* 529, 336–343. doi: 10.1038/nature17042

Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M., and Lund, O. (2016). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data.

J. Antimicrob. Chemother. 71, 2484–2488. doi: 10.1093/jac/dkw184

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi: 10.1038/s41587-020-0439-x

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., et al. (2019). Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* 63:e00483-19. doi: 10.1128/AAC.00483-19

Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., et al. (2017). ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* 3:e000131. doi: 10.1099/mgen.0.000131

Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 6:90. doi: 10.1186/s13073-014-0090-6

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

Lapidus, A. L., and Korobeynikov, A. I. (2021). Metagenomic data assembly – the way of decoding unknown microorganisms. *Front. Microbiol.* 12:613791. doi: 10.3389/fmicb.2021.613791

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Maguire, F., Jia, B., Gray, K. L., Lau, W. Y. V., Beiko, R. G., and Brinkman, F. S. L. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb. Genom.* 6:mgen000436. doi: 10.1099/mgen.0.000436

McArthur, A. G., and Wright, G. D. (2015). Bioinformatics of antimicrobial resistance in the age of molecular epidemiology. *Curr. Opin. Microbiol.* 27, 45–50. doi: 10.1016/j.mib.2015.07.004

Ng, C., Tay, M., Tan, B., Le, T. H., Haller, L., Chen, H., et al. (2017). Characterization of metagenomes in urban aquatic compartments reveals high prevalence of clinically relevant antibiotic resistance genes in wastewaters. *Front. Microbiol.* 8:2200. doi: 10.3389/fmicb.2017.02200

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

Panunzi, L. G. (2020). sraX: a novel comprehensive resistome analysis tool. *Front. Microbiol.* 11:52. doi: 10.3389/fmicb.2020.00052

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes *de novo* assembler. *Curr. Protoc. Bioinform.* 70:e102. doi: 10.1002/cpbi.102

Shlemov, A., and Korobeynikov, A. (2019). "PathRacer: racing profile HMM paths on assembly graph," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds I. Holmes, C. Martín-Vide, and M. Vega-Rodríguez (AlCoB, Algorithms for Computational Biology. Springer Verlag), 11488, 80–94. doi: 10.1007/978-3-030-18174-1_6

Silva, G. J., Correia, M., Vital, C., Ribeiro, G., Sousa, J. C., LeitÃ£o, R., et al. (2002). Molecular characterization of bla IMP-5, a new integron-borne metallo-Î^2-lactamase gene from an *Acinetobacter baumannii* nosocomial isolate in Portugal. *FEMS Microbiol. Lett.* 215, 33–39. doi: 10.1111/j.1574-6968.2002.tb11366.x

Smith, S. D., Colgan, P., Yang, F., Rieke, E. L., Soupir, M. L., Moorman, T. B., et al. (2019). Investigating the dispersal of antibiotic resistance associated genes from manure application to soil and drainage waters in simulated agricultural farmland systems. *PLoS ONE* 14:e0222470. doi: 10.1371/JOURNAL.PONE.0222470

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988

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.

ORIGINAL RESEARCH
published: 03 December 2021
doi: 10.3389/fmicb.2021.766206

MDRSA: A Web Based-Tool for Rapid Identification of Multidrug Resistant *Staphylococcus aureus* Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry

Chia-Ru Chung[1†], Zhuo Wang[2†], Jing-Mei Weng[1], Hsin-Yao Wang[3,4], Li-Ching Wu[5], Yi-Ju Tseng[3,6], Chun-Hsien Chen[3,7], Jang-Jih Lu[3,8,9], Jorng-Tzong Horng[1,10]* and Tzong-Yi Lee[2]**

[1] Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, [2] School of Life and Health Sciences, Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China, [3] Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan, [4] Ph.D. Program in Biomedical Engineering, Chang Gung University, Taoyuan, Taiwan, [5] Department of Biomedical Sciences and Engineering, National Central University, Taoyuan, Taiwan, [6] Department of Information Management, National Central University, Taoyuan, Taiwan, [7] Department of Information Management, Chang Gung University, Taoyuan, Taiwan, [8] College of Medicine, Chang Gung University, Taoyuan, Taiwan, [9] Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan, Taiwan, [10] Department of Bioinformatics and Medical Engineering, Asia University, Taichung City, Taiwan

As antibiotics resistance on superbugs has risen, more and more studies have focused on developing rapid antibiotics susceptibility tests (AST). Meanwhile, identification of multiple antibiotics resistance on *Staphylococcus aureus* provides instant information which can assist clinicians in administrating the appropriate prescriptions. In recent years, matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has emerged as a powerful tool in clinical microbiology laboratories for the rapid identification of bacterial species. Yet, lack of study devoted on providing efficient methods to deal with the MS shifting problem, not to mention to providing tools incorporating the MALDI-TOF MS for the clinical use which deliver the instant administration of antibiotics to the clinicians. In this study, we developed a web tool, MDRSA, for the rapid identification of oxacillin-, clindamycin-, and erythromycin-resistant *Staphylococcus aureus*. Specifically, the kernel density estimation (KDE) was adopted to deal with the peak shifting problem, which is critical to analyze mass spectra data, and machine learning methods, including decision trees, random forests, and support vector machines, which were used to construct the classifiers to identify the antibiotic resistance. The areas under the receiver operating the characteristic curve attained 0.8 on the internal (10-fold cross validation) and external (independent testing) validation.

Edited by:
*Liang Wang,
Xuzhou Medical University, China*

Reviewed by:
*Costas C. Papagiannitsis,
University of Thessaly, Greece
Matthew Leon Faron,
Medical College of Wisconsin,
United States*

***Correspondence:**
*Jang-Jih Lu
janglu45@gmail.com
Jorng-Tzong Horng
horng@db.csie.ncu.edu.tw
Tzong-Yi Lee
leetzongyi@cuhk.edu.cn*

†These authors have contributed equally to this work

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

Received: *28 August 2021*
Accepted: *28 October 2021*
Published: *03 December 2021*

Citation:
*Chung C-R, Wang Z, Weng J-M, Wang H-Y, Wu L-C, Tseng Y-J, Chen C-H, Lu J-J, Horng J-T and Lee T-Y (2021) MDRSA: A Web Based-Tool for Rapid Identification of Multidrug Resistant Staphylococcus aureus Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry.
Front. Microbiol. 12:766206.
doi: 10.3389/fmicb.2021.766206*

The promising results can provide more confidence to apply these prediction models in the real world. Briefly, this study provides a web-based tool to provide rapid predictions for the resistance of antibiotics on *Staphylococcus aureus* based on the MALDI-TOF MS data. The web tool is available at: http://fdblab.csie.ncu.edu.tw/mdrsa/.

Keywords: antibiotics susceptibility test, multidrug resistance, MALDI-TOF MS, machine learning, AST (antibiotic susceptibility testing)

INTRODUCTION

Over the past few decades, inappropriate use of antibiotics has brought out the growth of antibiotic resistance (ABR). More specifically, ABR is the ability of a bacterium to resist the effects of a treated drug and leads to the drug's ineffectiveness. Using alternative drugs or higher doses of antibiotics to defeat ABR is one of the solutions. However, overusing, underusing, or even misusing the drugs accelerates the growth of ABR. Additionally, it could lead to a bacterium being resistant to a variety of antibiotics, which is knowns as multidrug resistance (MDR), or even called "superbugs." Meanwhile, superbugs are a huge threat to global health today. One of the well-known superbugs is methicillin-resistant *Staphylococcus aureus* (MRSA), which has become a severe issue all over the world (Wolters et al., 2011; Clerc et al., 2014; Mather et al., 2016).

Staphylococcus aureus, a Gram-positive bacterium, is a microorganism commonly found on the skin. These carriers are not symptomatic. However, the pathogen occasionally causes severe diseases including skin, wounds, urinary tract, lung infections, bacteremia, and food poisoning (Naber, 2009). Antibiotics can effectively cure most *Staphylococcus aureus* infections, but MRSA is a bacterium that can resist methicillin and other antibiotics such as oxacillin (OX), penicillin, amoxicillin, and cephalosporin, which are improperly used and produce resistance. It is widely believed that the incorrect use of antibiotics is one of the causes of drug resistance. MRSA has a variety of antibiotic resistance and is generally considered a nosocomial pathogen which causes high mortality (Noskin et al., 2005). Therefore, it is very important to rapidly distinguish between methicillin-sensitive *Staphylococcus aureus* (MSSA) and MRSA.

There are several steps in the current process for determining the treatment of infectious diseases in clinical microbiology. When the doctor suspects that the patient is suffering from a certain infectious disease, the specimens of the infected site are collected for testing. After the specimen collection is completed, the bacterial culture is adopted to provide further bacterial identification. While confirming the bacteria, several antibiotic susceptibility tests (AST) are performed to decide the treatment. In general, it takes about 2–3 days to culture the bacteria and obtain the AST results (Lowy, 2003). Although the standard experiments are highly accurate, the time cost is also high. Before obtaining the AST reports, it is highly dependent on the physicians' experience to treat patients. Yet, empirical treatments might inadvertently cause more serious drug resistance. In short, the rapid information of AST can reduce ineffective use of drugs.

With the rapid development of antibiotic resistance, several methods for rapid identification of antibiotic resistance have been proposed, such as polymerase chain reaction (PCR) assays and, more recently, the matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS). MALDI-TOF MS is a proteomic tool that measures the molecules including proteins or peptides in the sample. The peptides that are associated with antibiotic resistance might be detected through MALDI-TOF mass spectra. Although qPCR, RT-qPCR, ddPCR, and modified 16S sequencing which obtain AST information in only a few hours could attain high performance, MALDI-TOF MS has more potential to become a convenient and efficient method for identification of antibiotic resistance. The primary reason is that MALDI-TOF MS has already been routinely used in many clinical microbiology laboratories, and there is no additional cost for those that have a MALDI-TOF MS. The mass spectra, generated by MALDI-TOF MS, are composed of peaks of specific mass—to—charge ratios (M/Z) with different intensities, which correspond to a reproducible fingerprint of a certain microorganism (Wang et al., 1998). Consequently, a number of studies have investigated the performance of MALDI-TOF MS on identification of bacterial strains (Ryzhov and Fenselau, 2001; Bizzini et al., 2010; Wang et al., 2018, 2019), and further explored the antibiotics resistance to bacteria (Singhal et al., 2015; Vrioni et al., 2018). Meanwhile, several studies have reported the significant effect on clinical microbiology (Psaroulaki and Chochlakis, 2018; Vrioni et al., 2018; Angeletti and Ciccozzi, 2019; Rodríguez-Sánchez et al., 2019; Welker et al., 2019). In brief, recognizing the pattern of the peptides would serve as a fingerprint for identifying antibiotic resistance in the study, and hence using MALDI-TOF MS to realize the rapid AST in clinical microbiology is promising.

According to the large amount of AST reports collected by Chang Gung Memorial Hospital, the percentages of resistant to erythromycin (E) and clindamycin (CC) were about 50%, which can be seen in **Supplementary Figure 1**. This implies that providing instant information about the use of them is as critical as the identification of MRSA. However, none of the studies used substantial data or provided a web-based prediction tool for the rapid identifications of oxacillin-, clindamycin-, and erythromycin-resistant *Staphylococcus aureus*. Therefore, the major purpose of this study is to develop a web-based prediction tool, MDRSA, for the rapid identification of multiple drugs resistant to *Staphylococcus aureus* based on a significant amount of MALDI-TOF MS data. Clinicians would obtain instant guidelines about the use of antibiotics for the *Staphylococcus aureus* infection. Additionally, the analysis for the informative peaks would provide more indications for the resistance. In short, development of rapid identification models does contribute an impact on the clinical management of patients with infectious diseases.

MATERIALS AND METHODS

Bacterial Isolates

A total of 20,212 and 5,005 clinical isolates were collected from two medical centers (CGMH Linkou branch and CGMH Kaohsiung branch). These two centers are around 330 km apart. Both centers serve as the referral centers in the regions. It should be noted that these data were collected from the current routine process for determining the treatment of infectious diseases in clinical microbiology. All clinical specimens were collected from all the wards continuously. The specimen types included blood, respiratory tract specimen (sputum, bronchial wash, and bronchoalveolar lavage), sterile cavity fluid (ascites, pleural effusion, pericardial effusion, cerebrospinal fluid, and synovial fluid), urine, and wound. Note that the data collected from Linkou and Kaohsiung branches were regarded as a training set and an independent set, respectively. All the processes of identifying *Staphylococcus aureus* and its resistance strictly followed the Clinical and Laboratory Standard Institute (CLSI) guidelines. **Table 1** shows the amount of data in training and independent testing sets. More than 81% of the isolates were recovered from the patient's sputum, pus, wounds, and blood specimens as shown in **Supplementary Table 1**.

Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry Data Acquisition

MALDI-TOF MS was used to identify the bacterial species and was conducted on Microflex LT (Bruker Daltonik GmbH, Bremen, Germany) benchtop instrument. All isolates were identified as *Staphylococcus aureus* by Bruker MALDI-TOF MS, and the measurement procedures were following the manufacturer's instructions (Bruker Daltonik GmbH, Bremen, Germany). Mass spectra were acquired in a linear positive mode within a range of +2 kV to +20 Kv and the nitrogen laser frequency was set as 60 Hz.

The species of *Staphylococcus aureus* was analyzed and reported on Biotyper 3.1 software (Bruker Daltonics). Biotyper provided the intensity and the signal quality of the peaks. For each isolate, the maximum number of peaks was set up to 200 and the acceptable quality is larger 2.0 which is the benchmark from the instruction of Biotyper 3.1. Furthermore, by using Flexanalysis 3.4 (Bruker Daltonik GmbH, Bremen, Germany) we could get a mass list, the parameters were set as follows: centroid peak detection algorithm for peak finding; Top Hat method for baseline subtraction; signal-to-noise threshold was set as 2; the minimum peak width expected in the spectrum was set as 6 M/Z; the maximal number of peaks was set as 200; relative intensity threshold was set as 0%; minimum intensity threshold was set as 0, and height was set as 80%. In this investigation, spectra ranging from 2,000 to 20,000 M/Z were acquired for further analysis.

TABLE 1 | Number of data in training and independent testing sets.

	Training set		Independent testing set	
Antibiotics	**Resistant (%)**	**Susceptible (%)**	**Resistant (%)**	**Susceptible (%)**
Oxacillin	10,735 (53.11)	9,477 (46.89)	2,399 (47.93)	2,606 (52.07)
Clindamycin	9,297 (46.00)	10,915 (54.00)	1,880 (37.56)	3,125 (62.44)
Erythromycin	11,304 (55.93)	8,908 (44.07)	2,584 (51.63)	2,421 (48.37)

Spectral Data Processing

Even in the same experimental steps and environment, the MALDI-TOF mass spectra of the same isolates may still be different. Specifically, the strong peaks on different MALDI-TOF mass spectra of the same strain may not be located at the same M/Z (Lin et al., 2005; AlMasoud et al., 2014), and we called this problem a shifting problem. Consequently, preprocessing for each single mass spectrum before constructing the models is an essential step, especially for large-scale data derived from the clinical medicine.

In order to deal with the peak shifting problem that appears in MALDI-TOF MS data, the kernel density estimation (KDE) was adopted to estimate the actual location of the peaks. Specifically, KDE is a non-parametric method to estimate the probability density function (PDF) of a random variable (Sheather and Jones, 1991). The M/Z values were regarded as the random variable. We then applied the KDE with Gaussian kernel to estimate the PDF of the M/Z values, which can be represented as

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) = \frac{1}{\sqrt{2\pi}nh}\sum_{i=1}^{n}\exp\{-\frac{1}{2}(\frac{x - x_i}{h})^2\} \tag{1}$$

where $x_1, x_2, \ldots, x_n$ are all M/Z values derived from all spectra, h is the bandwidth, which is also the smoothing parameter, n is the number of M/Z values, and K is the kernel function.

To obtain M/Z patterns for resistant and susceptible spectra, we used the function "stats.gaussian_kde," provided by SciPy (Virtanen et al., 2020), to estimate their PDFs in this study. It should be noted that the bandwidth is a critical parameter for employing KDE. The parameter "bw_method" provided in "stats.gaussian_kde" can be used to determine it. More specifically, if "bw_method" is a scalar, the bandwidth will be the scalar multiplied by the standard deviation of the sample. After the PDFs of M/Z patterns for resistant and susceptible spectra were obtained, the local modes derived from two PDFs were retrieved and concatenated to be a one spectrum with several peaks. Then we removed the duplicate values to construct a reference spectrum template. In addition to removing the duplicate values, the distance between two adjacent local modes less than three were also removed. The minimum width of two adjacent peaks expected in a spectrum was set as 6 M/Z in Flexanalysis 3.4. Finally, these M/Z values formed the final reference spectrum template. **Figure 1** demonstrates the flow chart of constructing a reference spectrum template. Note that 0.0006, 0.0008, 0.001, 0.0012, and 0.0014 were the values of "bw_method" in this study and used to generate different reference spectra. Their corresponding bandwidths were 2.08, 2.77, 3.46, 4.15, and 4.84. The peaks in every spectrum were then aligned to the nearest ones in the reference spectrum accordingly. **Supplementary Figure 2** illustrates the alignment for mass spectrum of Isolate A. More specifically, the purple line is the PDF of the population, so all the peaks of a mass spectrum should be shifted to the nearest benchmark which is the local maximum of PDF. **Supplementary Table 2** all features (peaks) used for developing oxacillin (OX), clindamycin (CC), and erythromycin (E) models.

The amount of the intensity may be influenced by various factors like temperature, instrument set-up, storage, and manual operation (Baumann et al., 2005), so we need to process the raw spectra data first. In this study, we scaled each intensity by its spectrum's maximum intensity. The definition of the formula is given below:

$$y_{ij}^{*} = \frac{y_{ij}}{\max\{y_{ij}|i = 1,\ 2,\ \ldots,\ n_j\}} \tag{2}$$

where y_{ij}^{*} and y_{ij} are the scaled and original intensities for the *j*th spectrum at the *i*th peak, respectively, and max ($y_{ij}|\ i = 1, 2, \ldots, n_j$) is the maximum intensity for the *j*th spectrum which contains n_j peaks.

Model Construction

After preprocessing the MS data, we adopted three machine learning (ML) algorithms, including decision tree (DT), random forest (RF), and support vector machine (SVM), to build up the classification models to predict the antibiotic resistance. Further information about the algorithms was then given in the next paragraph. The grid search with 10-fold cross validation was implemented on the training set for each bandwidth. When the optimal parameters were obtained, the independent testing set was used to evaluate the performance based on the model trained by the whole training set.

DT is a commonly used method for building the classification models. DT is formed in a tree-like structure which is constructed by nodes and leaves. Each node represents a test on a feature and each branch stands for an outcome of the test. Lastly, each leaf represents the class resulting from all tests. The criterion for yielding the best classification is important. Classification and regression trees (CART) algorithm is one of the commonly used algorithms to produce the best classification. The CART algorithm is a greedy approach that allows each step to select an optimal feature to get the most information gain when selecting attributes (Breiman et al., 1984). The measurement for selecting the optimal feature is finding the minimum impurity. In this study, we used the Gini index as the approach for calculating the impurity, which is the most common assessment approach. For each selection, the sum of the Gini impurity for all branches will be calculated, and the minimum one will be the best selection. The function "sklearn.tree.DecisionTreeClassifier" in scikit-learn package was used to build the DT model (Pedregosa et al., 2011).

RF is another common machine learning classifier, composed of multiple optimized version of CARTs to build the prediction model. RF uses bootstrap aggregating (bagging), one of the ensemble learning methods, to make sure each tree randomly gets training sets and attributes. As illustrated in **Supplementary Figure 3**, the ensemble learning method trains multiple models and votes the result finally, and the data used in each model was randomly determined in reusable. The classification outcome of RF is determined by the mode of every individual tree output. Most of the time, compared to DT, RF performs well when dealing with many features. Other reasons we use RF are that the learning time is short, and it can assess the importance of features easily. In this research, the tool we used to build the RF classification model is the function "sklearn.ensemble.RandomForestClassifier" in scikit-learn package (Pedregosa et al., 2011).

Support vector machine (SVM) is another common supervised learning classification. SVM finds a hyperplane that can minimize the risk of misclassification. The method used to minimize the risk is to find a decision boundary that can maximize the boundaries between the two classes. As shown in **Supplementary Figure 4**, there are two classes on a plane. We can find many possible hyperplanes that can separate two classes, and the algorithm for SVM is to find the hyperplane that can "maximum" the distance (the largest margin) between two classes. In this study, we use the function "sklearn.linear_model.SGDClassifier" in scikit-learn package (Pedregosa et al., 2011).

Statistical Analysis

Chi-squared test and *t*-test were employed in this study to evaluate the capability of discriminating the resistance for an individual peak based on their presence and intensities, respectively. Specifically, the chi-squared test of independence was mainly conducted to test the correlations between two categorical variables. In short, the small *p*-values concluded that the presence of a specific peak was correlated to the resistance. On the other hand, *t*-test was used to compare the intensities between two groups. Similarly, the small *p*-value would refer to that the intensity of a specific peak was different between two groups.

Evaluation Metrics

In this study, we used accuracy (ACC), the area under the receiver operating characteristic curve (AUC), sensitivity (SN), very major error (VME), specificity (SP), major error (ME), and Matthew's correlation coefficient (MCC) as the performance measurements for our models. The definitions of these measurements are given below.

$$ACC = \frac{TP + TN}{TP + TN+FP + FN} \tag{3}$$

$$SN = \frac{TP}{TP + FN} \tag{4}$$

$$VME = \frac{FN}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{FP + TN} \tag{6}$$

$$ME = \frac{FP}{FP + TN} \tag{7}$$

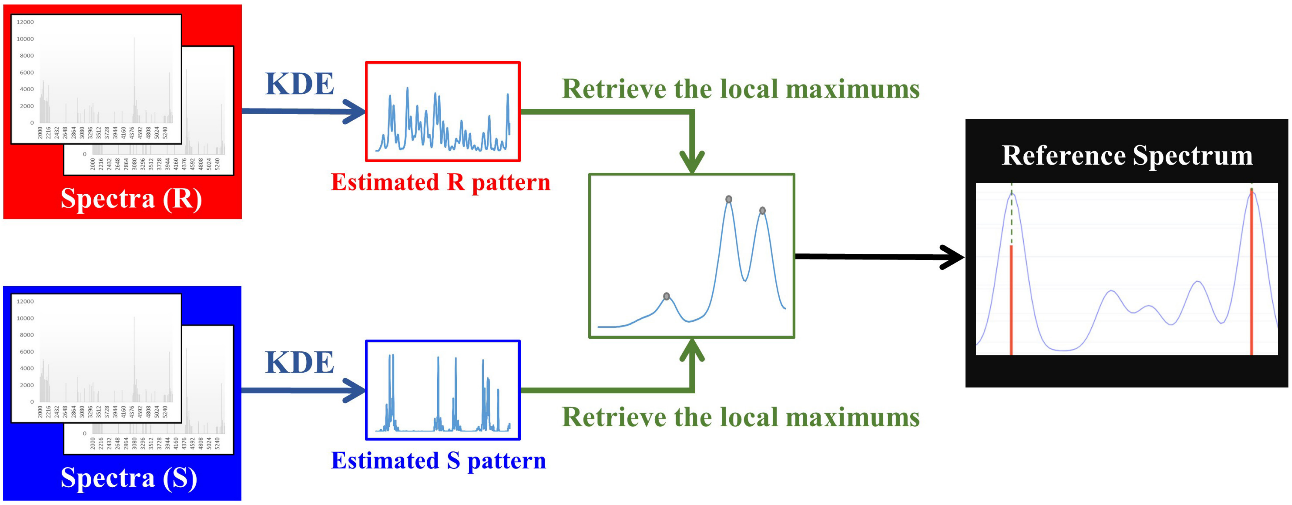


FIGURE 1 | Flow chart of constructing a reference spectrum template.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (8)$$

where TP is true positive which means the number of antibiotic-resistant isolates are correctly predicted by the classifier, TN is true negative which means the number of antibiotic-sensitive isolates are correctly predicted by the classifier, FP is false positive which means the number of antibiotic-sensitive isolates are wrongly predicted as antibiotic-resistant isolates by the classifier, and TN is false negative which means the number of antibiotic-resistant isolates are wrongly predicted as antibiotic-sensitive isolates by the classifier. Accuracy is the rate of the difference between the prediction results and the real results. MCC is the measurement to measure the quality of the binary classification. It returns a value between −1 and +1. If MCC returns +1, it means the prediction is perfect; if MCC is 0, if MCC returns −1, it represents the prediction is totally wrong. MCC considers the case that the sizes of the classes are very different and gives a balanced measurement.

In medicine, it is often determined by some thresholds whether the prediction result is true or false, and this threshold will affect the sensitivity and the specificity. In short, different threshold sets will lead to different prediction results. The distribution of the different threshold sensitivity and specificity can be plotted as the ROC curve, and the area under the ROC curve is called AUC. The most ideal case is AUC = 1, which is the case that the point locates on the upper left corner of the plot; when the AUC is 0.5, it represents a random selection of conditions, which means random guess. Most cases are within these two values. Through ROC and AUC, we can choose a more robust and stable model.

Development of a Web-Based Prediction Tool

We used hypertext markup language (HTML) and hypertext preprocessor (PHP) with python code to implement a web-based prediction tool in the backend upon submission of MALDI-TOF MS data. Each MS data should start with "BEGIN IONS" and end with "END IONS." This web-based prediction tool could predict one or more MS data for a submission. This web-prediction tool would list the prediction probabilities for the submitted MS data show the submitted MS figure with the important features.

RESULTS

MS Data Overview

Figure 2 shows the number of peaks in each spectrum according to different antibiotics resistance. Most of spectra preserved 50–150 peaks. Since most of the data were overlapped, there is no significant difference between the number of peaks between resistant and susceptible strains.

Figure 3 demonstrates the distribution of the number of spectra that were derived from oxacillin-, clindamycin-, and erythromycin-resistant/susceptible *Staphylococcus aureus* isolates at M/Z = 2,000–20,000. Since the range of M/Z is too large to obtain detailed information, we then further zoomed in to the M/Z range 2,000–3,000 to find some information (**Figure 4**). Peaks at M/Z = 2,360–2,500 are different between resistant and susceptible strains for all three antibiotics. We summed all intensities of resistant and susceptible isolates to observe the difference between them. It was still difficult to compare the difference between resistance and susceptibility (**Supplementary Figure 5**), so we zoomed in on these figures to find the differences. We found that it still has the difference of resistance and susceptibility at the range from 2,360 to 2,500 M/Z for oxacillin, clindamycin, and erythromycin (**Supplementary Figure 6**).

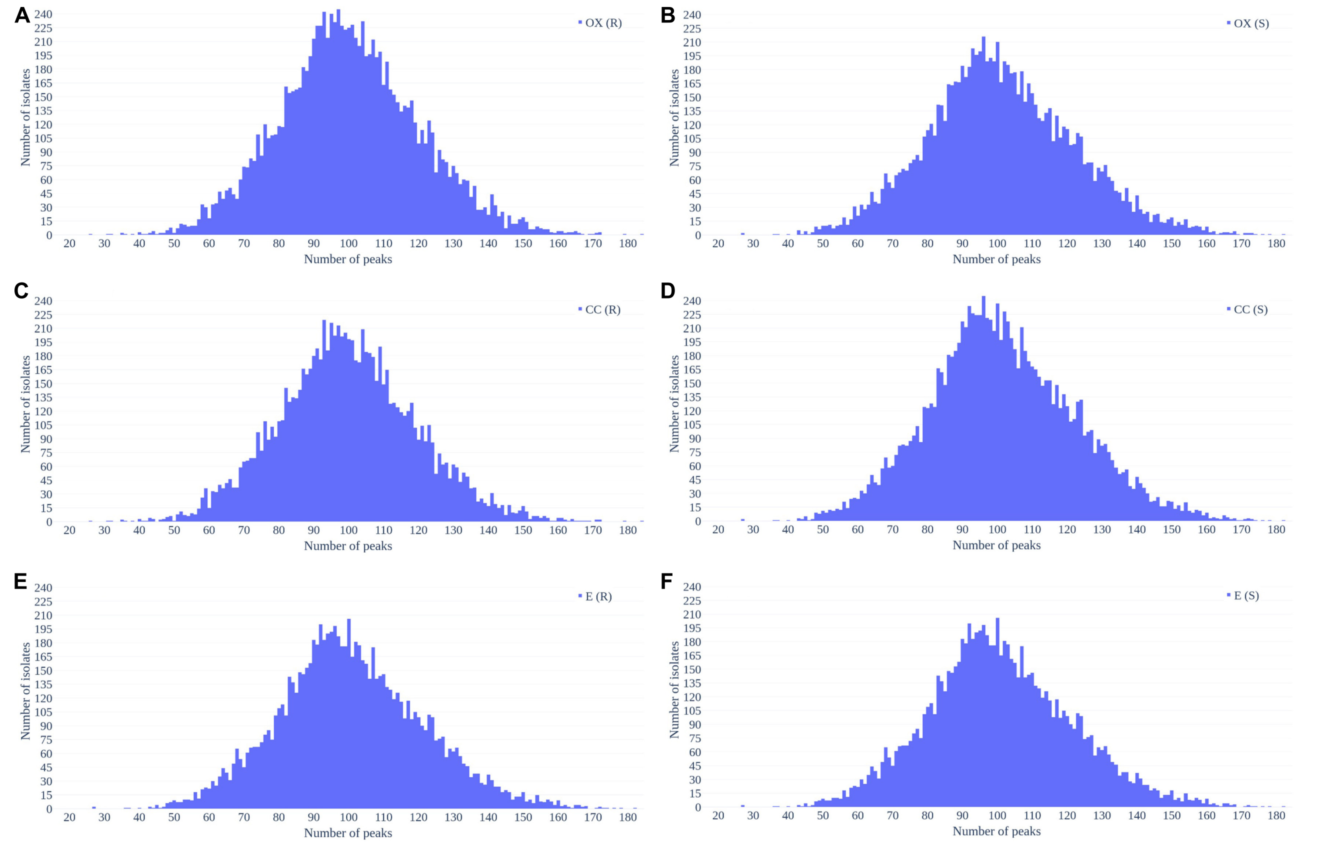


FIGURE 2 | Distribution of number of peaks retrieved from each spectrum in **(A)** oxacillin-resistant, **(B)** oxacillin-susceptible, **(C)** clindamycin-resistant, **(D)** clindamycin-susceptible, **(E)** erythromycin-resistant, and **(F)** erythromycin-susceptible *Staphylococcus aureus*.

Performance of Prediction Models

To build up a stable model, the parameters of the model are critical, especially the bandwidth. The bandwidth of Gaussian KDE is crucial. If the bandwidth is too large, the PDF will be too smooth; if the bandwidth is too small, it will be too harsh. We adopted 10-fold cross-validation models with different "bw_method" parameters and different ML algorithms to find the optimal bandwidth parameter. We tried the "bw_method" parameters from larger to smaller, and we found that when the "bw_method" parameter approaches 0.001, the accuracy tends to be stable. **Supplementary Table 3** shows the results of 10-fold cross-validation with the optimal parameters based on the grid search with different "bw_method" parameters of Gaussian KDE on the training set of oxacillin, clindamycin, and erythromycin, respectively. We could find that when the "bw_method" parameter was set to 0.0008, the standard deviations of oxacillin and clindamycin models were small and retained high accuracy Similarly, E model would reach the optimal when the bw_method is 0.001. Moreover, the highest accuracies are all built by the RF algorithm.

We adopted the optimal parameters to construct the RF-based models for the three antibiotics based on the whole training set. These models were then tested by the independent testing set and compared with those that did not use the KDE preprocessing. When the KDE preprocessing was adopted, the accuracies were 81.42, 82.20, and 74.63% for oxacillin, clindamycin, and erythromycin, respectively (**Table 2**). Comparing to the models that used data without KDE preprocessing, the accuracies derived from KDE were higher (6.04, 5.78, and 6.58%) for oxacillin, clindamycin, and erythromycin, respectively.

Forward Feature Selection

To obtain a more informative feature set, the feature importance scores calculated by RF was determined in this study. More specifically, we used the 70% training set to build up the classification models and calculated the features' importance scores based on the RF algorithm. The features were then ranked by their importance scores. After that, the feature was added in the model sequentially until the accuracy of the remaining 30% training set reached a plateau. **Supplementary Figure 7A** shows the trend of accuracy as the feature was added sequentially for the oxacillin model. When the number of features is 36, the model reaches a plateau and accuracy is 84.13%. The clindamycin model, demonstrated in **Supplementary Figure 7B**, attained a plateau at 37 features with an accuracy of 80.22%. Thirty-seven features were used to reach a plateau for the erythromycin model as shown in **Supplementary Figure 7C**. **Table 3** shows the performance of the selected features on the independent

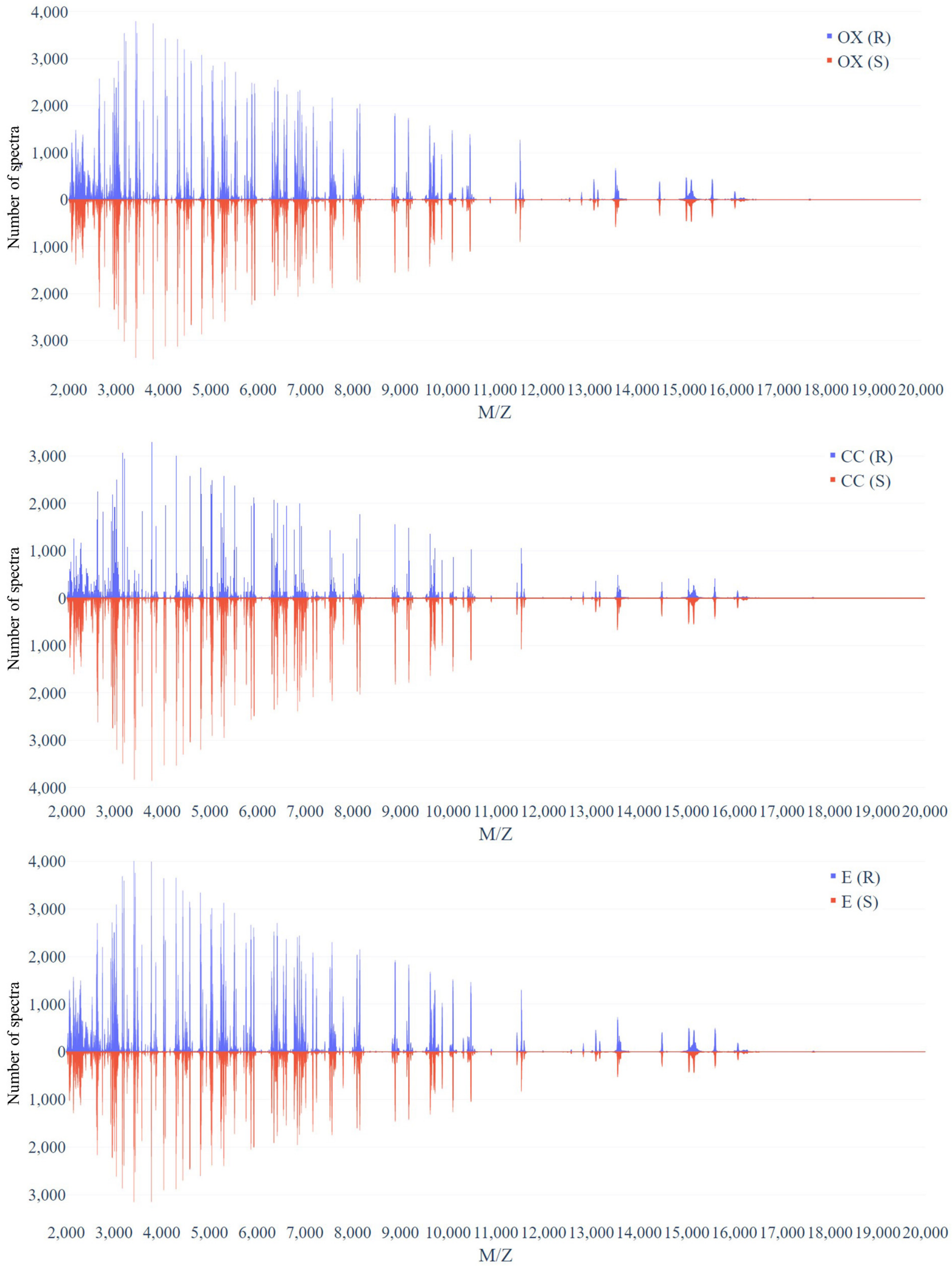


FIGURE 3 | Distribution of number of spectra that were derived from oxacillin- (upper), clindamycin- (middle), and erythromycin-resistant/susceptible (bottom) *Staphylococcus aureus* isolates at each M/Z.

testing set. When the number of features reduced to about 40, the accuracy was still around 80%. Furthermore, 589, 600, and 824 data were incorrectly called as sensitive for oxacillin, clindamycin, and erythromycin models, respectively. Meanwhile, 384, 280, and 442 data were incorrectly called as resistant for oxacillin, clindamycin, and erythromycin model, respectively.

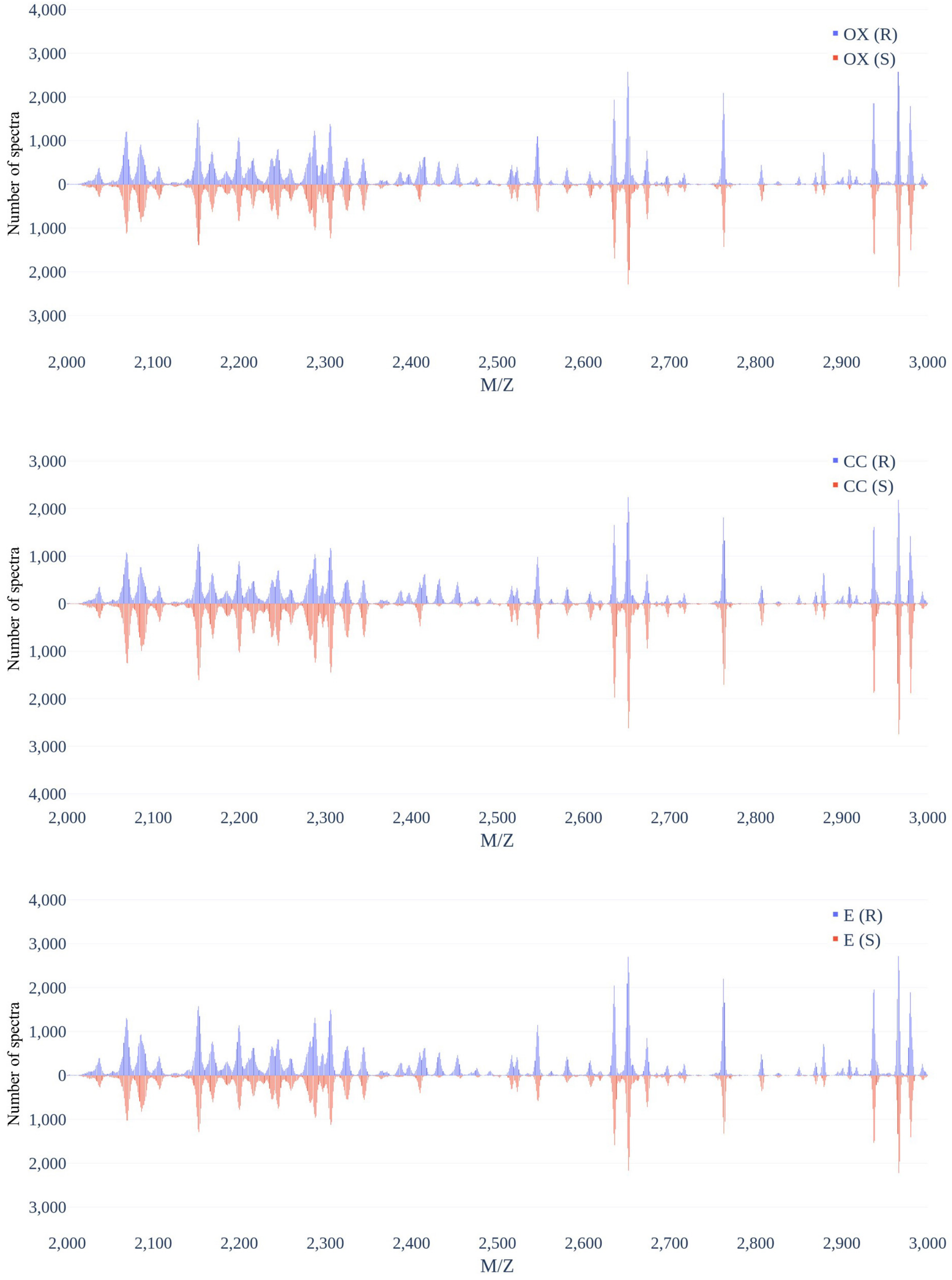


FIGURE 4 | Distribution of number of spectra that were derived from oxacillin- (upper), clindamycin- (middle), and erythromycin-resistant/susceptible (bottom) *Staphylococcus aureus* isolates at M/Z = 2,000–3,000.

Supplementary Table 4 lists all selected features for each model. We found most of the selected peaks were duplicated, but some peaks were selected uniquely for a certain model. More specifically, the peaks at 11,539, 4,526, and 3,297 M/Z were only selected by the oxacillin model. While the peaks at 2,910, 3,045, 2,966, and 7,568 M/Z were only included by the

TABLE 2 | Results of with or without kernel density estimation (KDE) preprocessing on independent testing set.

Antibiotics	Metrics	Without KDE preprocessing	Using KDE preprocessing
OX	SN	0.7962	0.7524
	VME	0.2038	0.2476
	SP	0.7149	0.8711
	ME	0.2851	0.1289
	ACC	0.7538	0.8142
	AUC	0.7555	0.8117
CC	SN	0.7282	0.6489
	VME	0.2718	0.3511
	SP	0.7859	0.9261
	ME	0.2141	0.0739
	ACC	0.7642	0.8220
	AUC	0.7571	0.7875
E	SN	0.7693	0.6908
	VME	0.2307	0.3092
	SP	0.5857	0.8055
	ME	0.4143	0.1945
	ACC	0.6805	0.7463
	AUC	0.6775	0.7481

OX, oxacillin; CC, clindamycin; E, erythromycin; SN, sensitivity; VME, very major error; SP, specificity; ME, major error; ACC, accuracy; AUC, area under the receiver operating characteristic curve.

TABLE 3 | Performance of features selection on independent test set.

	Model		
	Oxacillin	Clindamycin	Erythromycin
Number of features	36	38	37
Sensitivity	0.7545	0.6809	0.6811
Very major error	0.2455	0.3191	0.3189
Specificity	0.8526	0.9104	0.8174
Major error	0.1474	0.0896	0.1826
Accuracy	0.8706	0.8242	0.7471
AUC	0.8036	0.7956	0.7493

AUC, Area under the receiver operating characteristic curve.

clindamycin model after the feature selection. The erythromycin model incorporated peaks at 6,524, 4,514, 5,004, and 2,652 M/Z, which were not selected by other models. In addition, the peak at 6,593 M/Z ranked first for the oxacillin and erythromycin models. But the clindamycin model ranked it at a 14th place. This implies that the characteristics of resistance to clindamycin would be different from oxacillin and erythromycin.

In order to further investigate the selected peaks, we used the chi-square test for comparing two proportions of the resistant and susceptible data. Additionally, we also employed the *t*-test for comparing the intensities for these two groups. The results of these two statistical tests are shown in **Supplementary Tables 5–7** for the oxacillin model, the clindamycin model, and the erythromycin model, respectively. In this study, the *p*-value less than 0.001 was claimed as statistically significant. Most *p*-values of chi-square tests for the selected peaks were shown the significant difference between resistant and susceptible data. Yet, some selected peaks did not indicate the statistical significance such as peaks at 6,553, 5,526, and 3,277 M/Z for the clindamycin model when the chi-square test was adopted (**Supplementary Table 5**). While the *t*-test was employed to compare two intensities, several peaks did not show the significant difference such as the peaks at 3,008, 3,045, 2,200, 6,424, 6,890, and 2,966 M/Z for the clindamycin model and the peaks at 6,553, 2,306, 3,056, 2,287, and 7,021 M/Z for the erythromycin model.

Figure 5 and **Supplementary Figures 8**, **9** demonstrate the top 9 selected peak distributions of the M/Z values without peak alignment for three models to further investigate the difference on oxacillin-, clindamycin-, and erythromycin-resistant/susceptible data, respectively. These figures also indicate that the resistant isolates have more chance to appear at some specific peaks than the susceptible ones such as peaks at 6,593, 2,414, 2,432, and 2,456 M/Z on oxacillin data; peaks at 2,414, 2,432, 2,456, and 7,595 M/Z on clindamycin data; and peaks at 6,593, 2,413, 2,432, and 2,456 M/Z on erythromycin data.

Investigation of Multidrug Resistance

A Venn diagram was used to demonstrate the multiple antibiotics resistance, which is shown in **Supplementary Figure 10**. About 41% (8,234/20,212) of isolates were resistant to three antibiotics, and 37% (7,455/20,212) of isolates were susceptible to three antibiotics. This implies that most of the isolates were either resistant to three antibiotics or susceptible to them. Due to the few numbers of only resistant to a specific antibiotic or two antibiotics, we constructed a binary classification model to discriminate that the isolate is resistant or susceptible to three antibiotics simultaneously. **Supplementary Table 8** shows the amount of data for this classification. Similarly, we used 10-fold cross validation to find the best parameters and models. The performance is shown in **Supplementary Table 9**. The best AUC obtained from the RF model had a bandwidth of 0.0006. According to the optimal parameters derived from the training set, the performances on the independent testing set were 0.7918 (sensitivity), 0.9053 (specificity), 0.8545 (accuracy), 0.7057 (MCC), and 0.8486 (AUC).

RDMDRSA Web Interface

Based on our method, an online prediction server—MDRSA—was developed to predict the possibility that an MS derived from a *Staphylococcus aureus* isolate might be resistant to a particular antibiotic. The best prediction models developed for identifying the oxacillin-, clindamycin-, and erythromycin-resistant were applied here. Screenshots of the website are shown in **Supplementary Figure 11**.

DISCUSSION AND CONCLUSION

In this study, we used the MALDI-TOF MS data from Chang Gung Memorial Hospital Linkou branch to build different ML models to identify the resistance of the *Staphylococcus aureus*, and the data from Chang Gung Memorial Hospital Kaohsiung branch were further adopted to evaluate these models.

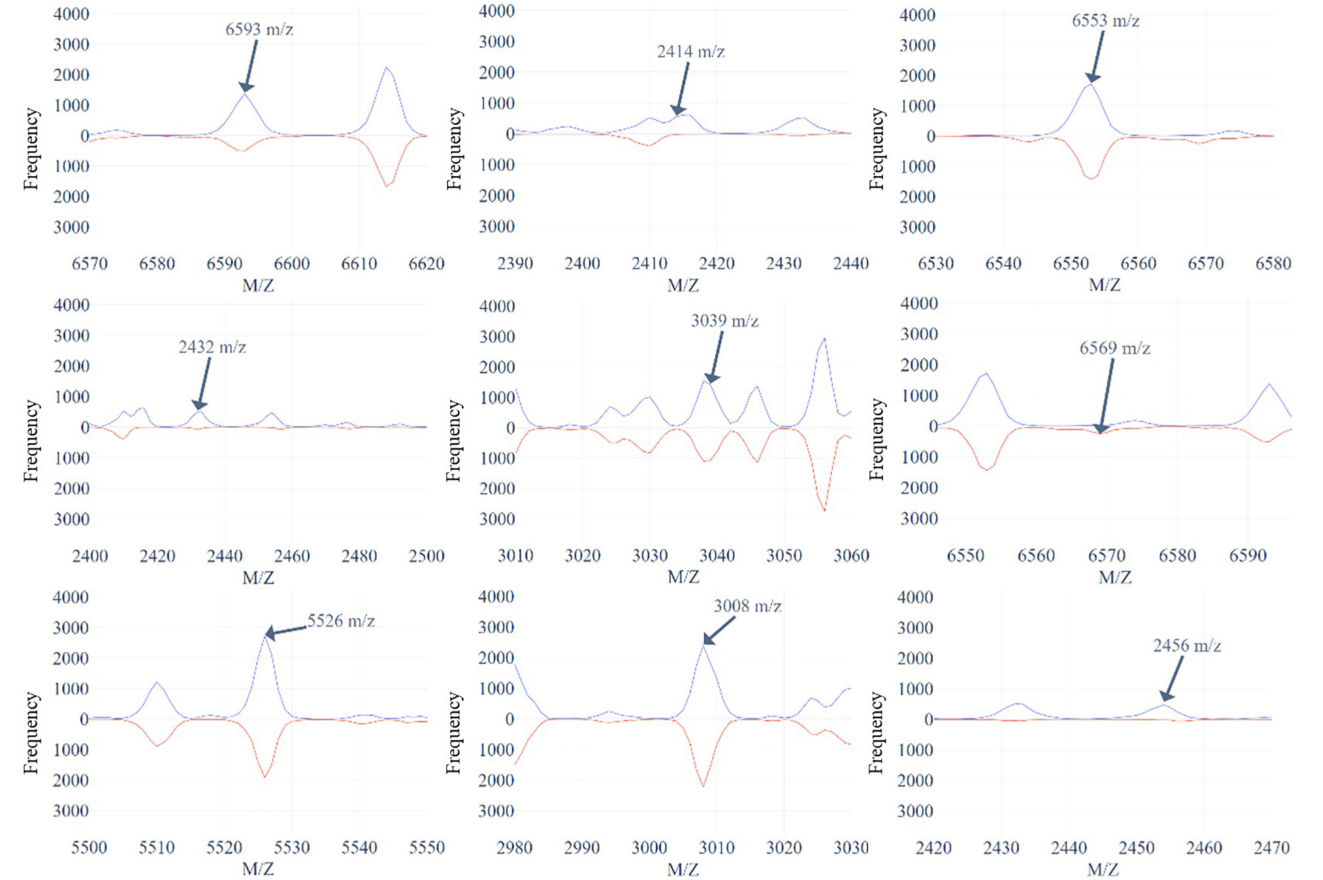


FIGURE 5 | The top 9 selected peaks distributions of the M/Z values without peaks alignment for oxacillin-resistant (red)/susceptible (blue) data.

Additionally, we adopted the Gaussian KDE method to deal with the shifting problem in MS data. Note that the bandwidth selection was based on the mean accuracy of the 10-fold cross validation. The accuracies of the 10-fold cross validation models were 86.28, 85.66, and 80.93% for oxacillin, clindamycin, and erythromycin models, respectively. Meanwhile, the accuracies of the independent testing were attained to 81.42, 82.20, and 74.63% for oxacillin, clindamycin, and erythromycin models, respectively. The forward feature selection was further used to reduce the dimension of features according to the order of importance derived from the RF. We then selected 36, 38, and 37 features for oxacillin, clindamycin, and erythromycin models, respectively. The accuracies of the models used selected features on the independent testing set were 80.56, 82.42, and 74.71% for oxacillin, clindamycin, and erythromycin models, respectively. The investigation of multiple drug resistance demonstrated that most isolates were either resistant to three antibiotics or susceptible to them. The accuracy of independent testing was 85.46% which was higher than the models that were used for identifying a specific resistance.

Previous studies were mainly devoted to identifying methicillin-resistant *Staphylococcus aureus* (MRSA), and figuring out their informative peaks (Wang et al., 2013, 2020; Josten et al., 2014; Østergaard et al., 2015; Camoez et al., 2016; Rhoads et al., 2016; Bai et al., 2017; Sogawa et al., 2017; Kim et al., 2019; Tang et al., 2019; Liu et al., 2021). Bai et al. (2017) proposed a genetic algorithm with a *t*-test based population seeding for wrapper feature selection on 727 *Staphylococcus aureus* clinical isolates' mass spectra derived from Vitek MS, and their accuracy based on support vector machine classifier was 0.72. Sogawa et al. (2017) utilized support vector machine to discriminate MRSA from methicillin-susceptible *Staphylococcus aureus* (MSSA) based on features derived from MALDI-TOF mass spectra. Their model reached prediction accuracies of over 85% and significantly reduced the time to initiation of targeted antibiotic treatment in comparison with phenotypic resistance profiling. Yet, they only considered 160 clinical isolates. Kim et al. (2019) developed discrimination models based on 320 clinical *Staphylococcus aureus* clinical isolates' mass spectra and 181 new ones were tested, and the DT had a sensitivity of 87.6%. Tang et al. (2019) applied different supervised ML models which are capable of distinguishing MRSA from MSSA. Even though their prediction accuracy was over 90%, only 20 isolates were used. Liu et al. (2021) used R to analyze 452 *Staphylococcus aureus* clinical isolates' mass spectra derived from Vitek MS, and the best area under the receiver operating characteristic curve was 0.89 by support vector machine. Compared with previous studies, our study used much clinical data and considered three antibiotics.

The limitation for analyzing antibiotic resistance through MALDI-TOF MS is that some antibiotic resistance-related peptides might not be detectable through mass spectra derived from MALDI-TOF MS when using the routine sample preparation protocol. This would limit the prediction for the antibiotic resistance. Yet, we incorporated data from two medical centers which are around 330 km apart. Given the spatial distribution of the two medical centers, we would detect the spectral pattern that is associated with antibiotic resistance. However, the possibility of detecting specific clones could not be fully excluded now without molecular strain typing data. Moreover, some factors including culture medium, bacteria lysis condition, and matrix crystallization condition, would have impact on the MALDI-TOF mass spectra and the subsequent identification of antibiotic resistance. Meanwhile, bacterial strains in different regions are quite diverse. Although it would be unsuitable to apply our models in other regions, we proposed a valid method to deal with the peak-shifting problem of MALDI-TOF MS. Specifically, the local MS data needed to be collected and our methods employed to develop the proper prediction models. On the other hand, our MALDI-TOF MS data were obtained from Bruker Daltonics GmbH. We did not compare with different MS data which was derived from different systems in the study. In addition, we did not further identify the proteins for the informative peaks. Even so, the results did show that the proportions of resistant were higher than the non-resistant ones for the selected peaks. The further identification of the informative peaks could provide a more comprehensive view on the mechanism of antibiotic resistance and would be valuable for the development of potential new treatments.

In this study, both accuracy and AUC for the internal (10-fold cross validation) and external (independent testing) validation attained 0.8. The promising results can provide more confidence to apply these prediction models in the real world. Briefly, this study provides a web-based tool to provide rapid predictions for the resistance of antibiotics on *Staphylococcus aureus* based on the MALDI-TOF MS data. In the future, a cross-national study is required. Given the high diversity of microorganisms across countries, it is not possible that the current prediction models can be used in other areas/countries without adjustment. Training and validating machine learning models based on locally relevant MALDI-TOF MS data are favorable.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

C-RC, J-MW, and H-YW carried out the data collection and curation. C-RC, ZW, and J-MW participated in the data analyses, model construction, and drafted the manuscript. C-RC, ZW, J-MW, H-YW, L-CW, and T-YL participated in the design of the study and performed the draft revision. J-TH, Y-JT, C-HC, T-YL, and J-JL conceived of the study, participated in its design and coordination, and helped to revise the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Ministry of Science and Technology, Taiwan (MOST108-2221-E-008-043-MY3 and MOST110-2320-B-182A-006-MY3) and Chang Gung Memorial Hospital (CMRPG3L0401 and CMRPG3L0431).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.766206/full#supplementary-material

REFERENCES

AlMasoud, N., Xu, Y., Nicolaou, N., and Goodacre, R. (2014). Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS) for the characterization of *Bacillus* and *Brevibacillus* species. *Anal. Chim. Acta* 840, 49–57. doi: 10.1016/j.aca.2014.06.032

Angeletti, S., and Ciccozzi, M. (2019). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: an updating review. *Infect. Genet. Evol.* 76:104063. doi: 10.1016/j.meegid.2019.104063

Bai, J., Fan, Z., Zhang, L., Xu, X., and Zhang, Z. (2017). "Classification of methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* using an improved genetic algorithm for feature selection based on mass spectra," in *Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology* (New York, NY: Association for Computing Machinery), 57–63.

Baumann, S., Ceglarek, U., Fiedler, G. M., Lembcke, J., Leichtle, A., and Thiery, J. (2005). Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin. Chem.* 51, 973–980. doi: 10.1373/clinchem.2004.047308

Bizzini, A., Durussel, C., Bille, J., Greub, G., and Prod'hom, G. (2010). Performance of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of bacterial strains routinely isolated in a clinical microbiology laboratory. *J. Clin. Microbiol.* 48, 1549–1554. doi: 10.1128/JCM.01794-09

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC press.

Camoez, M., Sierra, J. M., Dominguez, M. A., Ferrer-Navarro, M., Vila, J., and Roca, I. (2016). Automated categorization of methicillin-resistant *Staphylococcus aureus* clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. *Clin. Microbiol. Infect.* 22, 161.e1–161.e7. doi: 10.1016/j.cmi.2015.10.009

Clerc, O., Prod'hom, G., Senn, L., Jaton, K., Zanetti, G., Calandra, T., et al. (2014). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry and PCR-based rapid diagnosis of *Staphylococcus aureus* bacteraemia. *Clin. Microbiol. Infect.* 20, 355–360. doi: 10.1111/1469-0691.12329

Josten, M., Dischinger, J., Szekat, C., Reif, M., Al-Sabti, N., Sahl, H. G., et al. (2014). Identification of agr-positive methicillin-resistant *Staphylococcus aureus* harbouring the class A mec complex by MALDI-TOF mass spectrometry. *Int. J. Med. Microbiol.* 304, 1018–1023. doi: 10.1016/j.ijmm.2014.07.005

Kim, J. M., Kim, I., Chung, S. H., Chung, Y., Han, M., and Kim, J. S. (2019). Rapid discrimination of methicillin-resistant *Staphylococcus aureus* by MALDI-TOF MS. *Pathogens* 8:214. doi: 10.3390/pathogens8040214

Lin, S. M., Haney, R. P., Campa, M. J., Fitzgerald, M. C., and Patz, E. F. (2005). Characterising phase variations in MALDI-TOF data and correcting them by peak alignment. *Cancer Inform.* 1, 32–40.

Liu, X., Su, T., Hsu, Y.-M. S., Yu, H., Yang, H. S., Jiang, L., et al. (2021). Rapid identification and discrimination of methicillin-resistant *Staphylococcus aureus* strains via matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 35:e8972. doi: 10.1002/rcm.8972

Lowy, F. D. (2003). Antimicrobial resistance: the example of *Staphylococcus aureus*. *J. Clin. Invest.* 111, 1265–1273. doi: 10.1172/jci18535

Mather, C. A., Werth, B. J., Sivagnanam, S., Sengupta, D. J., and Butler-Wu, S. M. (2016). Rapid detection of vancomycin-intermediate *Staphylococcus aureus* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microbiol.* 54, 883–890. doi: 10.1128/JCM.02428-15

Naber, C. K. (2009). *Staphylococcus aureus* bacteremia: epidemiology, pathophysiology, and management strategies. *Clin. Infect. Dis.* 48, S231–S237. doi: 10.1086/598189

Noskin, G. A., Rubin, R. J., Schentag, J. J., Kluytmans, J., Hedblom, E. C., Smulders, M., et al. (2005). The burden of *Staphylococcus aureus* infections on hospitals in the United States: an analysis of the 2000 and 2001 Nationwide Inpatient Sample Database. *Arch. Intern. Med.* 165, 1756–1761. doi: 10.1001/archinte.165.15.1756

Østergaard, C., Hansen, S. G. K., and Møller, J. K. (2015). Rapid first-line discrimination of methicillin resistant *Staphylococcus aureus* strains using MALDI-TOF MS. *Int. J. Med. Microbiol.* 305, 838–847. doi: 10.1016/j.ijmm.2015.08.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Psaroulaki, A., and Chochlakis, D. (2018). Use of MALDI-TOF mass spectrometry in the battle against bacterial infectious diseases: recent achievements and future perspectives. *Expert Rev. Proteomics* 15, 537–539. doi: 10.1080/14789450.2018.1499469

Rhoads, D. D., Wang, H., Karichu, J., and Richter, S. S. (2016). The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci. *Diagn. Microbiol. Infect. Dis.* 86, 257–261. doi: 10.1016/j.diagmicrobio.2016.08.001

Rodríguez-Sánchez, B., Cercenado, E., Coste, A. T., and Greub, G. (2019). Review of the impact of MALDI-TOF MS in public health and hospital hygiene, 2018. *Eurosurveillance* 24:1800193. doi: 10.2807/1560-7917.ES.2019.24.4.1800193

Ryzhov, V., and Fenselau, C. (2001). Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Anal. Chem.* 73, 746–750. doi: 10.1021/ac0008791

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth selection method for Kernel density estimation. *J. R. Stat. Soc. Series B Stat. Methodol.* 53, 683–690.

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* 6:791. doi: 10.3389/fmicb.2015.00791

Sogawa, K., Watanabe, M., Ishige, T., Segawa, S., Miyabe, A., Murata, S., et al. (2017). Rapid discrimination between methicillin-sensitive and methicillin-resistant *Staphylococcus aureus* using MALDI-TOF mass spectrometry. *Biocontrol Sci.* 22, 163–169. doi: 10.4265/bio.22.163

Tang, W., Ranganathan, N., Shahrezaei, V., and Larrouy-Maumus, G. (2019). MALDI-TOF mass spectrometry on intact bacteria combined with a refined analysis framework allows accurate classification of MSSA and MRSA. *PLoS One* 14:e0218951. doi: 10.1371/journal.pone.0218951

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272.

Vrioni, G., Tsiamis, C., Oikonomidis, G., Theodoridou, K., Kapsimali, V., and Tsakris, A. (2018). MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance: current achievements and future perspectives. *Ann. Transl. Med.* 6:240. doi: 10.21037/atm.2018.06.28

Wang, H.-Y., Chung, C.-R., Wang, Z., Li, S., Chu, B.-Y., Horng, J.-T., et al. (2020). A large-scale investigation and identification of methicillin-resistant *Staphylococcus aureus* based on peaks binning of matrix-assisted laser desorption ionization-time of flight MS spectra. *Brief. Bioinformatics* 22:bbaa138. doi: 10.1093/bib/bbaa138

Wang, H. Y., Lee, T. Y., Tseng, Y. J., Liu, T. P., Huang, K. Y., Chang, Y. T., et al. (2018). A new scheme for strain typing of methicillin-resistant *Staphylococcus aureus* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. *PLoS One* 13:e0194289. doi: 10.1371/journal.pone.0194289

Wang, H. Y., Li, W. C., Huang, K. Y., Chung, C. R., Horng, J. T., Hsu, J. F., et al. (2019). Rapid classification of group B Streptococcus serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques. *BMC Bioinformatics* 20:703. doi: 10.1186/s12859-019-3282-7

Wang, Y. R., Chen, Q., Cui, S. H., and Li, F. Q. (2013). Characterization of *Staphylococcus aureus* isolated from clinical specimens by matrix assisted laser desorption/ionization time-of-flight mass spectrometry. *Biomed. Environ. Sci.* 26, 430–436. doi: 10.3967/0895-3988.2013.06.003

Wang, Z., Russon, L., Li, L., Roser, D. C., and Long, S. R. (1998). Investigation of spectral reproducibility in direct analysis of bacteria proteins by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 12, 456–464. doi: 10.1002/(sici)1097-0231(19980430)12:8<456::aid-rcm177>3.0.co;2-u

Welker, M., Van Belkum, A., Girard, V., Charrier, J.-P., and Pincus, D. (2019). An update on the routine application of MALDI-TOF MS in clinical microbiology. *Expert Rev. Proteomics* 16, 695–710.

Wolters, M., Rohde, H., Maier, T., Belmar-Campos, C., Franke, G., Scherpe, S., et al. (2011). MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. *Int. J. Med. Microbiol.* 301, 64–68. doi: 10.1016/j.ijmm.2010.06.002

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

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 February 2022
doi: 10.3389/fmicb.2021.770323

Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients

*Valeriia Ivanova[1], Ekaterina Chernevskaya[1,2], Petr Vasiluev[1,3], Artem Ivanov[4], Ivan Tolstoganov[5], Daria Shafranskaya[5], Vladimir Ulyantsev[4], Anton Korobeynikov[5], Sergey V. Razin[1,6], Natalia Beloborodova[2], Sergey V. Ulianov[1,6] and Alexander Tyakht[1,7]**

[1] Institute of Gene Biology Russian Academy of Sciences, Moscow, Russia, [2] Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia, [3] Research Centre for Medical Genetics, Moscow, Russia, [4] Computer Technologies Laboratory, ITMO University, Saint Petersburg, Russia, [5] Center for Algorithmic Biotechnologies, Saint Petersburg State University, Saint Petersburg, Russia, [6] Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [7] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology Russian Academy of Sciences, Moscow, Russia

OPEN ACCESS

Edited by:
Liang Wang,
Xuzhou Medical University, China

Reviewed by:
Martial Marbouty,
UMR 3525 Génétique des Génomes, France
Teresa Nogueira,
Instituto Nacional Investigaciao Agraria e Veterinaria (INIAV), Portugal

Correspondence:
Alexander Tyakht
a.tyakht@gmail.com

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

Received: *03 September 2021*
Accepted: *25 November 2021*
Published: *03 February 2022*

Citation:
Ivanova V, Chernevskaya E, Vasiluev P, Ivanov A, Tolstoganov I, Shafranskaya D, Ulyantsev V, Korobeynikov A, Razin SV, Beloborodova N, Ulianov SV and Tyakht A (2022) Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients. Front. Microbiol. 12:770323. doi: 10.3389/fmicb.2021.770323

Gut microbiome in critically ill patients shows profound dysbiosis. The most vulnerable is the subgroup of chronically critically ill (CCI) patients – those suffering from long-term dependence on support systems in intensive care units. It is important to investigate their microbiome as a potential reservoir of opportunistic taxa causing co-infections and a morbidity factor. We explored dynamics of microbiome composition in the CCI patients by combining "shotgun" metagenomics with chromosome conformation capture (Hi-C). Stool samples were collected at 2 time points from 2 patients with severe brain injury with different outcomes within a 1–2-week interval. The metagenome-assembled genomes (MAGs) were reconstructed based on the Hi-C data using a novel hicSPAdes method (along with the bin3c method for comparison), as well as independently of the Hi-C using MetaBAT2. The resistomes of the samples were derived using a novel assembly graph-based approach. Links of bacteria to antibiotic resistance genes, plasmids and viruses were analyzed using Hi-C-based networks. The gut community structure was enriched in opportunistic microorganisms. The binning using hicSPAdes was superior to the conventional WGS-based binning as well as to the bin3c in terms of the number, completeness and contamination of the reconstructed MAGs. Using *Klebsiella pneumoniae* as an example, we showed how chromosome conformation capture can aid comparative genomic analysis of clinically important pathogens. Diverse associations of resistome with antimicrobial therapy from the level of assembly graphs to gene content were discovered. Analysis of Hi-C networks suggested multiple "host-plasmid" and "host-phage" links. Hi-C metagenomics is a promising technique for investigating clinical microbiome samples. It provides a community composition profile with increased details on bacterial gene content and mobile genetic elements compared to conventional metagenomics. The ability of Hi-C binning to encompass the MAG's plasmid content facilitates metagenomic evaluation of virulence and drug resistance dynamics in clinically relevant opportunistic pathogens. These findings will help to identify the targets for developing cost-effective and rapid tests for assessing microbiome-related health risks.

Keywords: Hi-C metagenomics, critical care, antibiotic resistance, plasmids, *Klebsiella*, metagenome-assembled genome, binning, gut microbiome

INTRODUCTION

The number of patients with long-term dependence on support systems in the ICU is growing with the improvement in the quality of medical care. Chronically critically ill (CCI) patients are a heterogeneous group of patients after long stay in ICU characterized by the unique physiology including hormonal and metabolic changes with hypermetabolic and hypercatabolic state (Nelson et al., 2010; Parfenov et al., 2020), cognitive impairment, myopathy, inflammation (Cox, 2012) and increased susceptibility to infection (Nierman and Nelson, 2002). Gastrointestinal tract diseases can be observed in most such patients due to impaired swallowing, insufficient physical activity, horizontal position, feeding through a gastric tube or stoma, fluid and electrolyte disorders. The profound dysbiosis of the gut microbiome has a great influence on the gastrointestinal tract (Foster, 2016), leading to inflammation and tissue injury. These changes form a vicious pathological circle that prevents recovery. In addition, these patients receive multiple classes of antibiotics during hospitalization to treat ventilator-associated pneumonia, bloodstream infections caused by use of a central venous catheter, as well as the urinary tract infections, thus contributing toward the evolution of their microbiome into a reservoir of virulent multidrug-resistant nosocomial pathogens. Previously, we characterized the taxonomic composition of microbiome in CCI patients using 16S rRNA sequencing (Chernevskaya et al., 2020). The observed dysbiosis was linked to prognosis and reflected in the altered spectrum of microbially produced phenolic metabolites in blood and stool (Chernevskaya et al., 2021). It suggested the necessity of further investigation of the genetic potential of the microbial species abundant in CCI – including mobile genetic elements – using more powerful approaches like "shotgun" metagenomics.

Recently, conventional metagenomics have been combined with chromosome conformation capture techniques like Hi-C and 3C-seq to enable a deeper exploration of complex microbial communities. Besides the environmental microbiome (Baudry et al., 2019; Stalder et al., 2019), such techniques have been applied to mammal host-associated communities (Marbouty et al., 2017; Stewart et al., 2018; Bickhart et al., 2022). In the human microbiome field, all but one Hi-C metagenomic survey (Kent et al., 2020) performed to date investigated the gut community of healthy subjects (Press et al., 2017; DeMaere et al., 2020; Marbouty et al., 2021). In these studies, the overimposement of the paired Hi-C reads reflecting the information about chromosome spatial proximity onto the metagenomic assembly allowed better binning of the contigs into metagenome-assembled genomes (MAGs). Few software tools developed for Hi-C genome deconvolution have been published (Baudry et al., 2019; DeMaere and Darling, 2019).

By exploiting the fact that the chromosome interaction signal is higher across the genomic sequences present in the same microbial cell, it was possible to suggest specific "phage-microbe" and "plasmid-microbe" links. The mobile genetic elements are responsible for the horizontal gene transfer (HGT) that is of high biomedical relevance due to the transmission of antibiotic resistance and virulence factors genes. Therefore, application of Hi-C metagenomics to the human microbiome in the clinical context is promising for assessing microbiome-associated health risks in patients, especially in the immunocompromised ones. The only study of the alterations of gut microbiome in disease focused on the neutropenic patients undergoing hematopoietic stem cell transplantation (Kent et al., 2020).

In our study, we established an experimental and bioinformatic pipeline for Hi-C metagenomic analysis involving novel algorithms and applied it to explore the functional dynamics of gut dysbiosis in a pilot set of samples from CCI patients, with particular focus on evaluating such essential advantages of the technique compared to the conventional WGS as improved reconstruction of microbial genomes and a possibility to link antibiotic resistance genes, plasmids and viruses to their hosts.

MATERIALS AND METHODS

Study Design

This prospective observational study was performed in the Department of Intensive Care at the Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russian Federation. The stool samples were collected from two CCI patients: patient A – a 75 year old female after an intracerebral hemorrhage and patient B – a 74 year old male after an ischemic stroke. Both patients were on prolonged mechanical ventilation and enteral tube feeding (high calorie, low-residue) and received antibiotics (**Figure 1**). At the first time point, each had a suspected bacterial infection (pathogens were isolated from trachea, chest CT scans showed pneumonia). The second time point at which the stool was collected was day 7 for patient A (with negative clinical dynamics) and day 14 – for patient B (positive clinical dynamics). The

Abbreviations: ARG, antibiotic resistance gene; CCI, chronically critically ill; HGT, horizontal gene transfer; ICU, intensive care unit; MAG, metagenome-assembled genome; SCG, single-copy core genes; UTIs, urogenital tract infections; VF, virulence factors.

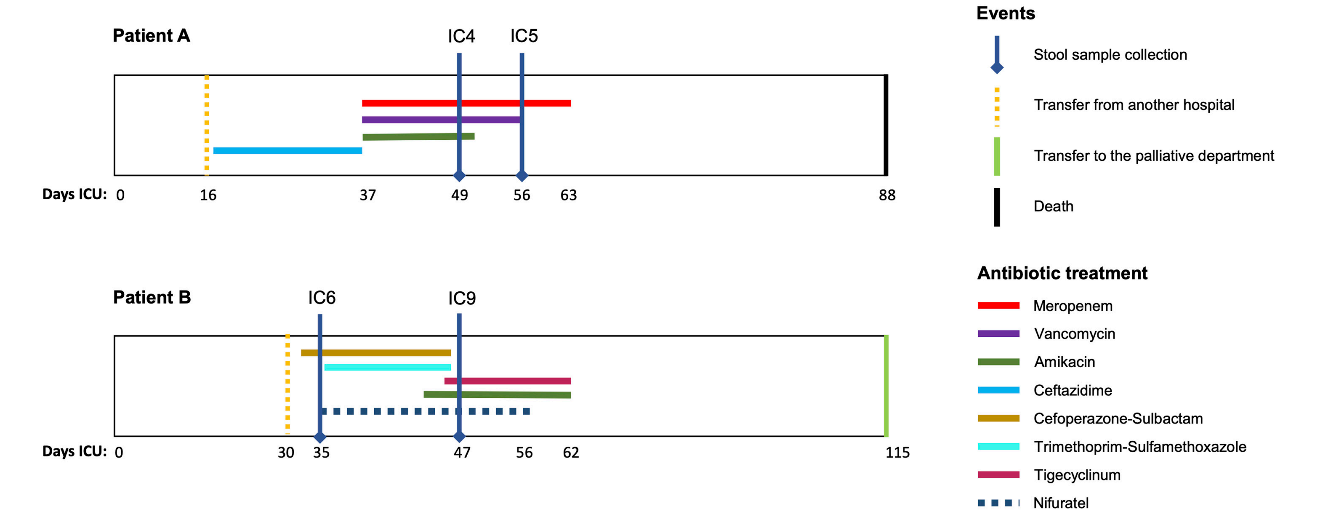


FIGURE 1 | Antimicrobial therapy timelines for the patients. The time courses of the two patients **(A,B)** included in the study are shown along with the periods (days) of antimicrobial drug administration (colored lines). Vertical lines indicate the key time points for the patients.

detailed clinical data is provided in the Additional File 1: **Supplementary Table 1**.

Sample Preparation and Sequencing

The WGS (metagenomic) libraries were prepared using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina exactly according to the manufacturer's instructions. Total genomic DNA was isolated as follows. Cell material was incubated in a 1× TE buffer at 65°C for 14–16 h in the presence of proteinase K (1 μg/μl) and 0.5% of SDS. DNA was then purified by single phenol-chloroform extraction followed by ethanol precipitation with 20 μg/ml glycogen (Thermo Fisher Scientific) as the co-precipitator. After precipitation, the pellets were dissolved in 50 μl 10 mM Tris–HCl pH 8.0. To remove residual RNA, samples were treated with 25 μg of RNase A (Thermo Fisher Scientific) for 45 min at 37°C. To remove residual salts and DTT, the DNA was additionally purified using Agencourt AMPure XP beads (Beckman Coulter). Then 20–100 ng of the purified DNA was used for WGS library preparation.

The Hi-C libraries were prepared as described below, in two replicates per sample. The sample was resuspended in saline solution (NaCl 0.9%), homogenized and centrifuged at 200 × *g* for 5 min, to precipitate the debris. Supernatant was removed in a separate tube and centrifuged at 10,000 × *g* for 5 min. The pellet was homogenized for 40 s in tubes containing Lysing Matrix A (MP Biomedicals) and a 6 mm ceramic sphere using an MP Biomedicals FastPrep-24 instrument at 6 m/s, then was resuspended in 1 ml of saline solution and centrifuged at 10,000 × *g* for 5 min 3 more times. Then the pellet was resuspended in 1 ml of fixing solution (NaCl 0.9%, formaldehyde 3%) and incubated for 20 min at 22°C with tube inversion every 2 min. The reaction was stopped by the addition of 2 M glycine to give a final concentration of 125 mM. Cells were centrifuged (17,000 × *g*, 10 min, 4°C), resuspended in 50 μl of 1 × PBS, snap-frozen in liquid nitrogen and stored at −80°C. Defrozen cells were mechanically disrupted using a Dounce homogenizer and additionally lysed in 1 ml isotonic buffer [50 mM Tris–HCl pH 8.0, 150 mM NaCl, 0.5% (v/v) NP-40 substitute (Fluka), 1% (v/v) Triton-X100 (Sigma), 1 × Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific)] on ice for 15 min. Cells were centrifuged at 20,000 × *g* for 5 min at 4°C, resuspended in 200 μl of 1 × NEBuffer 2 (NEB), and pelleted again. The pellet was resuspended in 200 μl of 0.3% SDS in 1 × NEBuffer 2 and incubated at 37°C for 1 h. Then, cells were centrifuged at 20,000 × *g* for 5 min at 4°C, washed with 200 μl of 1 × NEBuffer 2 and resuspended in 1 × CutSmart buffer (NEB) supplemented with 1% of Triton X-100 (Sigma). 100 U of *Hpa*II enzyme (NEB) were added, and the DNA was digested overnight (14–16 h) at 37°C with shaking (1,400 rpm). On the following day, additional 100 U of *Hpa*II enzyme were added, and the cells were incubated for an additional 2 h. *Hpa*II was then inactivated by incubation at 65°C for 20 min. After *Hpa*II inactivation, the cells were harvested for 10 min at 20,000 × *g*, washed with 300 μl of 1 × T4 DNA ligase buffer (Fermentas), and resuspended in 300 μl of 1 × T4 DNA ligase buffer. Cohesive DNA ends were ligated in the presence of 75 U of T4 DNA ligase (Fermentas) at 16°C for 4 h. The cross-links were reversed by overnight incubation at 65°C in the presence of proteinase K (1 μg/μl) (Sigma) and 0.5% of SDS. After cross-link reversal, the DNA was purified by single phenol-chloroform extraction followed by ethanol precipitation with 20 μg/ml glycogen (Thermo Fisher Scientific) as the co-precipitator. After precipitation, the pellets were dissolved in 100 μl 10 mM Tris–HCl pH 8.0. To remove residual RNA, samples were treated with 50 μg of RNase A (Thermo Fisher Scientific) for 45 min at 37°C. To remove residual salts and DTT, the DNA was additionally purified using Agencourt AMPure XP beads (Beckman Coulter). The DNA was then dissolved in 500 μl of sonication buffer [50 mM Tris–HCl (pH 8.0), 10 mM EDTA, 0.1% SDS] and sheared to a size of approximately 100–1,000 bp using a VirSonic 100 (VerTis). The samples were concentrated (and simultaneously purified) using AMICON Ultra Centrifugal Filter Units to a total volume of approximately 50 μl. The DNA ends were repaired by adding 62.5 μl MQ water, 14 μl of 10 × T4 DNA ligase reaction buffer (Fermentas), 3.5 μl of 10 mM dNTP mix (Fermentas), 5 μl of 3 U/μl T4 DNA polymerase (NEB), 5 μl of 10 U/μl T4 polynucleotide kinase (NEB), 1 μl of 5 U/μl Klenow DNA polymerase (NEB), and then incubating at 20°C for 30 min. The DNA was purified with Agencourt AMPure XP beads and eluted with 127 μl of 10 mM Tris–HCl (pH 8.0). To perform an A-tailing reaction, the DNA samples were supplemented with 15 μl 10 × NEBuffer 2, 3 μl of 10 mM dATP, and 4.5 μl of 5 U/μl Klenow (exo-) (NEB). The reactions were carried out for 30 min at 37°C in a PCR machine, and the enzyme was then heat-inactivated by incubation at 65°C for 20 min. The DNA was purified using Agencourt AMPure XP beads and eluted with 100 μl of 10 mM Tris–HCl (pH 8.0). Illumina TruSeq adapters were ligated by adding 12 μl 10 × T4 DNA ligase reaction buffer (Fermentas), 6 μl of Illumina TruSeq adapters and 2 μl of 5 U/μl T4 DNA ligase (Fermentas). Adapter ligation was performed at 22°C overnight. DNA was then purified with Agencourt AMPure XP beads and eluted with 30 μl of 10 mM Tris–HCl (pH 8.0). Test PCR reactions containing 5 μl of the samples were performed to determine the optimal number of PCR cycles required to generate sufficient PCR products for sequencing. The PCR reactions were performed using KAPA High Fidelity DNA Polymerase (KAPA) and Illumina PE1.0 and PE2.0 PCR primers (10 pmol each). The temperature profile was 5 min at 98°C, followed by 6, 9, 12, 15, and 18 cycles of 20 s at 98°C, 15 s at 65°C, and 20 s at 72°C. The PCR reactions were separated on a 2% agarose gel containing ethidium bromide, and the number of PCR cycles necessary to obtain a sufficient amount of DNA was determined based on the visual inspection of gels (typically 10–12 cycles). Four preparative PCR reactions were performed for each sample. The PCR mixtures were combined, and the DNA was purified using Agencourt AMPure XP beads and eluted with 50 μl of 10 mM Tris–HCl (pH 8.0).

The sequencing of the WGS and Hi-C libraries was performed on the Illumina HiSeq platform in 2 × 150 bp reads format.

Additionally, the abundance of selected gut taxa was measured using multiplex real-time PCR with fluorescent detection with the Colonoflor-16 kit (Alfalab, Russia).

Analysis of the WGS and Hi-C Data

We established the analytical pipeline for combined analysis of the "shotgun" readsets and their chromosome conformation capture counterparts; the workflow diagram is outlined in Additional File 1: **Supplementary Figure 1**. At the WGS prefiltering step, read merging and adapter removal was performed by bbmerge v.37.62 (with default parameters and k = 61, adapter = default). The WGS reads were assembled using SPADES v.3.15 (Prjibelski et al., 2020) in "meta" mode. Replicates of Hi-C libraries were pooled before processing and then filtered with BBMap (bbduk). For each sample, the binning of contigs was performed without the Hi-C data – into WGS-MAGs, *via* MetaBat 2 (Kang et al., 2019) – and using Hi-C data – into Hi-C MAGs, *via* the novel hicSPAdes algorithm.[1] As an additional option, the Hi-C MAGs have been produced using the previously published bin3c algorithm (DeMaere and Darling, 2019).

Taxonomic profiling of metagenomic reads was performed using MetaPhlAn2 (Truong et al., 2015) and MiCoP (LaPierre et al., 2019). The MAG quality was evaluated using CheckM (Parks et al., 2015). The taxonomy of each MAG was inferred using GTDB-Tk (Chaumeil et al., 2019). Circular packing plots of MAGs were generated using packcircles R package.[2] The quality of Hi-C libraries was evaluated with qc3C (DeMaere and Darling, 2021). Comparative genome analysis and visualization of the MAGs and published reference genomes was performed using anvi'o pipeline (Eren et al., 2015) in the pangenomic workflow including: 'anvi-script-FASTA-to-contigs-db' script – for contigs database construction for each MAG, 'anvi-run-ncbi-cogs' – for annotating the genes according to the NCBI Clusters of Orthologous Groups database, 'anvi-gen-genomes-storage' – for creating genome storage, 'anvi-pan-genome' – for generating pan database (particularly, search for amino acid sequence similarity with blastp and gene clusters identification), 'anvi-display-pan' – for visualization in the interactive interface, and 'anvi-summarize' – for generating static summary; the genes were predicted using Prodigal (Hyatt et al., 2010). Tree was constructed based on SCGs by FastTree and MAFFT, using 'anvi-gen-phylogenomic-tree' script. Genes encoding virulence factors were identified using the VFanalyzer tool with the VFDB database (Liu et al., 2019). UpSet plots were generated using the UpsetR package (Conway et al., 2017).

For evaluating the number of plasmid-like contigs in the assembly and the MAGs, PlasFlow (Krawczyk et al., 2018) was used (followed by blastn to the NCBI plasmid database for taxonomic validation). Prophage sequences in contigs were identified using PHASTER (Arndt et al., 2019). Search and annotation of antibiotic resistance genes (ARGs) in the MAGs and contigs were conducted with RGI (Resistance Gene Identifier)[3] using the CARD database (Alcock et al., 2020). The whole-metagenome resistome was assessed using the GraphAMR pipeline (Shafranskaya et al., 2021). GraphAMR overcomes the limitations of fragmented contigs in metagenomic assemblies (due to interspecies repeats, horizontal gene transfer and other mechanisms) *via* aligning the ARG profile Hidden Markov Model directly to the assembly graph with subsequent dereplication and identification. This approach allows for accurate and comprehensive recovery of ARGs without the necessity of their assembly into a complete contig. The method is able to detect an ARG even if it is not assembled (and thus spans multiple contigs).

For the construction of Hi-C contig networks, the contigs > 1,000 bp were selected as vertices. The Hi-C reads were mapped to the contigs using bwa. For each contig pair, the number of Hi-C read pairs connecting them was normalized using HiCzin[4] based on contigs' coverage and length as well as intra-species contacts. Intra-species contacts were estimated as

[1] https://cab.spbu.ru/software/hicspades/

[2] https://CRAN.R-project.org/package=packcircles

[3] https://card.mcmaster.ca/analyze/rgi

[4] https://github.com/dyxstat/HiCzin

the contacts between contigs within high-quality MAGs (with completeness > 80% and contamination < 5%). Using the normalized weights, we constructed distributions of intra-MAG and inter-MAG contacts intensity for the high-quality MAGs, for each sample (Additional File 1: **Supplementary Figure 2A**). Based on these histograms, we visually determined a threshold value of 0.6 – as it fits well across the samples to remove most inter-MAG contacts while retaining most intra-MAG contacts. Only the Hi-C links with normalized weight above this threshold were used to construct the network edges. During the analysis of "microbe-phage" links, we considered a viral contig to be connected to a bacterial MAG if it had a link to at least one of the MAG's contigs in the normalized interaction network with a weight above the threshold (0.6).

The classification of contigs as chromosomal/viral/plasmid was conducted using ViralVerify (Antipov et al., 2020) (with the -p flag used for the plasmid search). Draft taxonomic annotation of contigs was obtained using Kraken (Wood and Salzberg, 2014). The networks were visualized in Cytoscape (Su et al., 2014).

Taxonomic classification of the predicted viral contigs from assemblies was performed using DemoVir[5] to the levels of order and family. In an analysis complementary to the Hi-C-based approach, prediction of associations between viral contigs and bacterial MAGs was performed using VirMatcher – based on viral sequence matches to host CRISPR-spacers, integrated prophages in host genomes, host tRNA genes, and host k-mer signatures calculated by WisH (Gregory et al., 2020). Only the matches with final score ≥ 3 (according to the guidelines provided in the software repository)[6] were considered. Simulation of WGS reads (10 mln read pairs per sample) was performed using InSilicoSeq (Gourlé et al., 2019). For simulating Hi-C reads, the Sim3C tool (DeMaere and Darling, 2018) was used (5 mln read pairs per sample, read length 150 bp, *Hpa*II enzyme).

RESULTS

Basic Analysis of Gut Community Structure: Pronounced Dysbiosis

The WGS sequencing of 2 pairs of stool samples produced 109–131 mln read pairs per sample. A preliminary taxonomic profiling of the patients' gut metagenomes was performed using unique clade-specific gene markers (see section "Materials and Methods"). It revealed pronounced dysbiosis, particularly, with the levels of Proteobacteria 1–2 orders higher than observed from NGS microbiome surveys for the general Russian population (Tyakht et al., 2013; Klimenko et al., 2018; Volokh et al., 2019). The disruption of gut community structures has been confirmed *via* a complementary analysis using taxon-specific qPCR (Additional File 2: **Supplementary Table 2**). The decreased diversity was driven by *Bacteroidaceae* and various opportunist genera (*Klebsiella*, *Escherichia*, *Proteus*, *Bilophila*; see Additional File 1: **Supplementary Figure 3**). Besides the prokaryotes, there were fungal sequences normally not observed in healthy populations: the intracellular parasite *Enterocytozoon bieneusi* was omnipresent, with other detections including *Aspergillus niger* and *Candida glabrata* (Additional File 3: **Supplementary Table 3**).

Hi-C Allows to Obtain Higher Quality Metagenome-Assembled Genomes Compared to WGS

Next we investigated the microbiome composition of the patients at a deeper level *via* the reconstruction of MAGs (**Figure 2**). For each sample, the assembly was of relatively good quality including 91,381–139,339 contigs > 200 bp long, with a maximum length of 523,891–783,225 bp and the N50 value of 4,196–9,150 bp. According to the qc3c analysis, the estimated fraction of Hi-C reads was 18.28 – 47.55%, suggesting overall proper ligation.

Two types of MAGs were reconstructed for each sample – one being Hi-C-agnostic (WGS-MAGs) and another one exploiting the Hi-C linkage information (Hi-C MAGs). The conventional WGS binning was conducted using MetaBat2 as a state-of-art WGS binning algorithm. The Hi-C binning was performed in 2 versions. The first one used hicSPAdes – a novel binning and binning improvement tool that simultaneously exploits the information from Hi-C-derived links and topology of the assembly graph to improve the completeness and purity of MAG bins; the second – performed for comparison purposes – was the existing bin3c algorithm (see Additional File 4: **Supplementary Table 4** for summary statistics of libraries, assemblies and binnings).

Even when the Hi-C binning was performed using bin3c (Additional File 5: **Supplementary Tables 5A–D**), the number of produced high-quality MAGs (completeness > 80%, contamination < 5%) was higher than for the WGS (bin3c: 16–25 vs. MetaBAT2: 11–20) (Additional File 6: **Supplementary Table 6**). The contamination across the high-quality MAGs did not exceed 11% for Hi-C, while some WGS-MAGs had levels up to 300% or higher (**Figure 3**).

The superiority of the Hi-C approach was even more pronounced when the novel hicSPAdes was used for obtaining the Hi-C MAGs (23–27 high-quality MAGs and contamination < 7% for all MAGs) (Additional File 7: **Supplementary Table 7**). Compared to the WGS, the completeness was significantly higher for the Hi-C MAGs produced by hicSPAdes (while no significance was achieved in the case of bin3c; *t*-test for MAGs pooled across the 4 samples, $p = 0$ and 0.58, respectively).

High contamination levels close to multiples of 100% sporadically manifested by some WGS-MAGs were due to erroneous conglomeration of 2 or more genomes. The Hi-C approach allowed to resolve such cases. As an illustrative example of this effect, a *Dysgonomonas* WGS-MAG from the sample IC6 had a 88.5% contamination. It corresponded to two high-quality Hi-C MAGs (*via* hicSPAdes) – each classified at genus level as *Dysgonomonas* with contamination < 1% and completeness of 98.01–99.93% (Additional Files 6, 7: **Supplementary Tables 6, 7**). The two Hi-C counterparts of a *Bacteroides dorei* WGS-MAG (contamination: 92.2%, strain

[5] https://github.com/feargalr/Demovir

[6] https://bitbucket.org/MAVERICLab/virmatcher

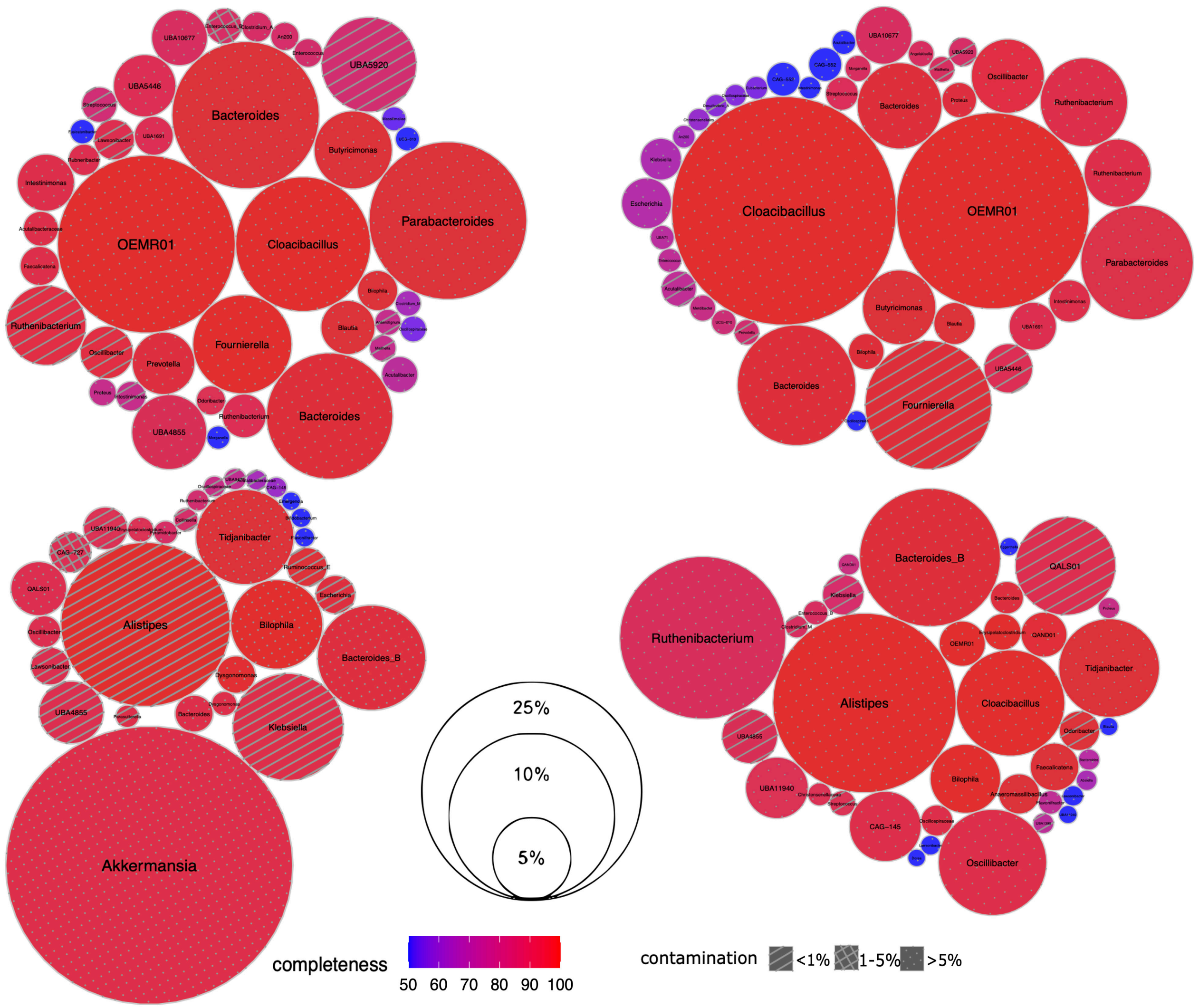


FIGURE 2 | Taxonomic composition of gut microbiome in CCI patients. For each sample, its set of recovered Hi-C MAGs (*via* hicSPAdes) is visualized as circle packing. Each circle represents a MAG labeled with genus-level taxonomy; MAG relative abundance (normalized by the total length of its contigs) is shown as its diameter, completeness – as color, and contamination – as fill pattern. Sample IDs from left to right, top to bottom: IC4, IC5, IC6 and IC9.

heterogeneity: 50.0%) were a *B. dorei* and *B. xylanisolvens*—with 0.38 and 0% contamination, respectively.

We analyzed the proportions of the assembly that were not binned into the MAGs. In terms of assembly length proportion, the Hi-C MAGs included 55.6–64.3% of the total contig length, while for the WGS, the sum was 54.9–63.4%. Considering a lower contamination in Hi-C MAGs, it suggests that they are generally more encompassing and provide more complete gene content for each member of the microbiome community, while balancing it with detailedness of species-level disentanglement.

As the set of abundant species was considerably overlapping between the timepoints, we also performed per-patient cross-assembly to assess how it improves the completeness of the reconstructed Hi-C MAGs (here we chose bin3c as an established binning algorithm to serve as a baseline). The results of the cross-assembly and comparison with the sample-wise Hi-C MAGs are shown in Additional File 5: **Supplementary Tables 5E,F**. In this analysis, we excluded the MAGs of low completeness (<10%) unclassified according to GTDBtk. The effect on MAG quality was ambiguous. For patient B, there were 16 (32.7%) cross-assembled MAGs that improved in quality (for some – dramatically) and 4 MAGs with novel taxonomy were obtained. For 20 MAGs, their completeness did not improve by remaining close to the maximum across the two time points; nine MAGs had their completeness decreased. For patient A, the respective numbers of Hi-C MAGs were: 25 (43.9%) – were improved, 10 novel, 11 – had similar completeness and 11 – decreased in completeness.

Noteworthy, the cross-assembly did not promote contamination for most Hi-C MAGs. Such an increase was observed for both patients: for patient A, only 4 of her MAGs with contamination < 5% received values > 5% in the

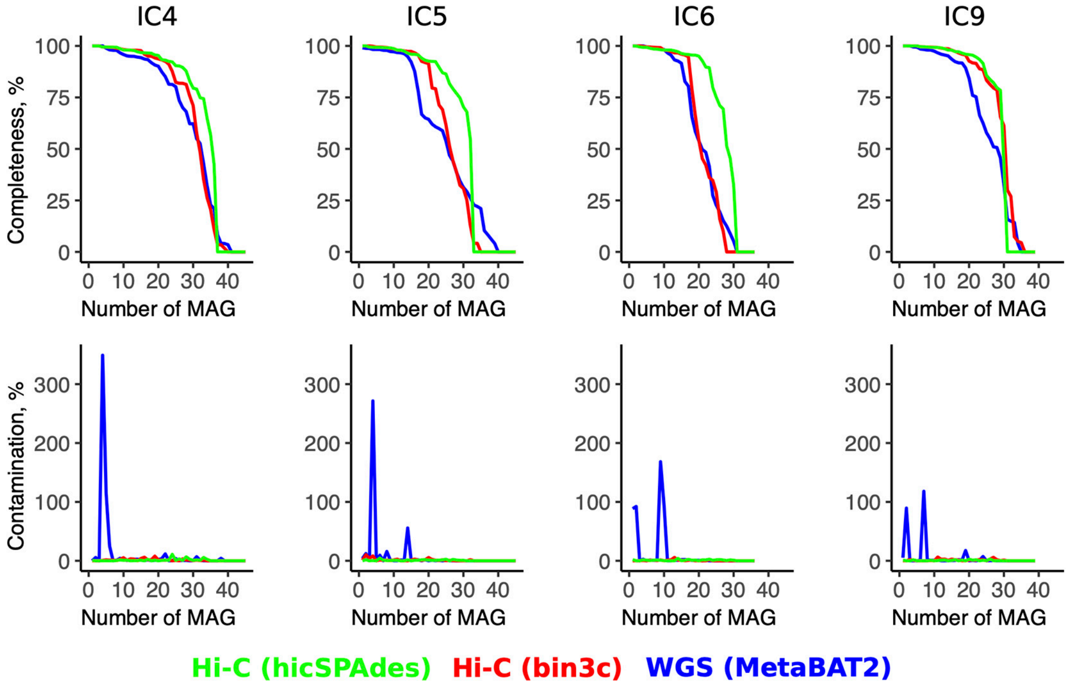


FIGURE 3 | Quality metrics for the Hi-C- and WGS-MAGs. The line plots show the values of MAGs' completeness (top row) and contamination (bottom row) for each sample (column-wise) according to the WGS binning and two versions of Hi-C binnings (shown in three colors). For each sample, the MAGs are sorted in the order of decreasing completeness.

cross-assembly (but remained below 15.2%); for patient B, there were two such cases (for them, contamination was < 11.5%).

Comparative Genome Analysis for Major Opportunist Taxa Facilitated by Hi-C

We evaluated the advantages of Hi-C MAGs to explore the virulence and drug resistance potential of the opportunist taxa enriched in the CCI microbiome. For the proof-of-principle, we selected the *Klebsiella pneumoniae* (Kp) – the species abundant in all 4 samples and represented by a single Hi-C MAG in each of them. (The hicSPAdes version of Hi-C MAGs were used in all samples but IC4 – in the latter, hicSPAdes did not produce a Kp MAG during the binning so we used the bin3c version instead).

After the gene prediction, the four MAGs were subject to clustering by their shared single-copy core genes (SCG) content similarity together with the reference genomes representative of the 3 known *K. pneumoniae* phylotypes (**Figure 4**). The results suggested that the CCI patients hosted *K. pneumoniae sensu stricto* (KpI phylotype), the one most commonly associated with human infection (Holt et al., 2015). The Kp MAGs clustered by subject; while the gene count was considerably lower for patient A than for B (5,547–5,262 vs. 5,759–6,403, according to anvi'o), the number of subject-wise genes persisting between the time points was, on the contrary, higher for A (2,097 vs. 1,407; **Figure 5**).

Evaluation of the Kp virulence potential from its Hi-C MAGs yielded 89–112 genes encoding virulence factors (VF; Additional File 8: **Supplementary Table 8**) suggesting these microbiomes host virulent Kp types. The VF lists included pili, fimbriae, efflux pumps, colibactin, capsule genes along with the RcsAB and RmpA systems regulating its production, the iron-scavenging siderophores salmochelin, aerobactin and yersiniabactin (the latter being the most common virulence factor associated with human *K. pneumoniae* infections); type VI secretion systems and the *rfb* locus responsible for lipopolysaccharide (LPS) biosynthesis. Allantoin utilization genes associated with liver hypervirulent strains were not detected in any of the MAGs. While the number of VF genes was close across all 4 samples, their proportion among all genes was higher for the samples from patient A (due to shorter Kp MAGs in both of them). Interestingly, only a few VF genes were subject-specific. For patient A, it was a type VI secretion system *tle1* phospholipase effector gene involved in bacterial competition. Another difference between the patients was in the set of fimbrial adherence determinants (likely acquired from *Salmonella*; all belonging to chaperone/usher fimbriae): in patient A, these were the genes of *stb* from γ4 clade, while in patient B – of *ste* and *stf* from π clade (Dufresne et al., 2018).

Noteworthy, the WGS-MAGs contained fewer VF genes than the Hi-C MAGs (WGS: 66–101; Hi-C: 89–112); their proportion was also lower – for all samples but IC6.

Plasmid Completeness of Metagenome-Assembled Genomes

Specifically in the case of *K. pneumoniae*, the conventional completeness and contamination metrics of its MAGs (as assessed *via* CheckM) were not considerably different between

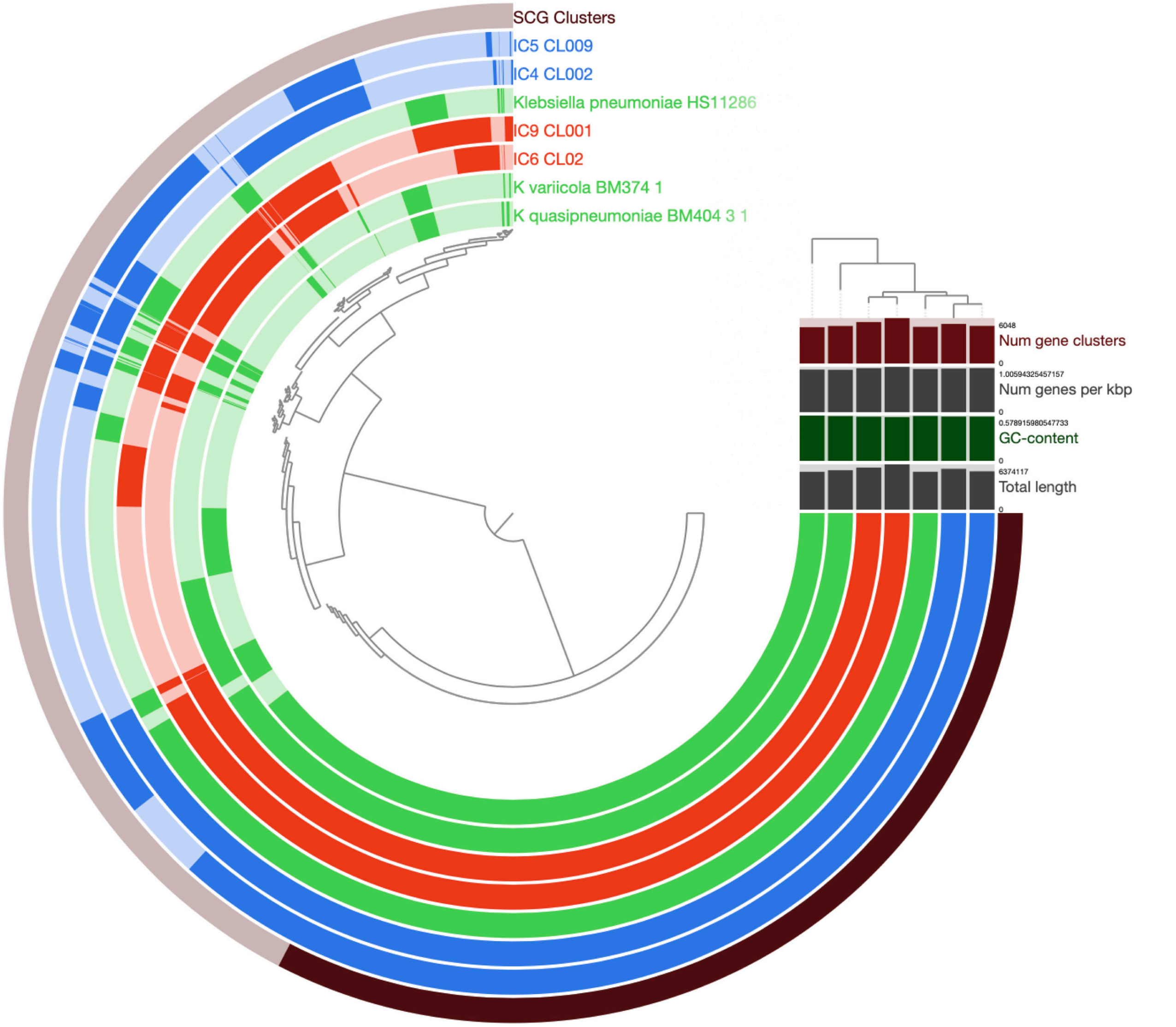


FIGURE 4 | Comparison of *K. pneumoniae* MAGs across the patients and time points. The genomes of *K. pneumoniae* strain HS11286, *K. variicola* BM374-1 and *K. quasipneumoniae* BM404-3-1 were included as external references. The circular hierarchical clustering diagram was constructed using the anvi'o pipeline (see section "Materials and Methods"). Each concentric circle represents a genome, while each radial ray corresponds to a gene (gene orthology cluster). The outermost circle shows the SCG genes (in black). The genomes/circles were hierarchically clustered based on their SCG sequence similarity, while the genes – on their prevalence pattern across the genomes. For the circles, the Kp MAGs of patient A and B are shown in blue and red, respectively; the reference genomes – in green. Saturated colors correspond to present genes, while pale ones – to the absent ones. In the upper right part, the total number of gene clusters, the gene density (number of genes per Kbp of genome), GC content (%) and total length are provided for each genome/MAG.

the WGS and Hi-C versions. The completeness was > 91% and contamination – < 3% in most cases. However, these metrics are not particularly focused on accounting for the accessory genes and plasmids. Plasmids are an important HGT channel responsible for dissemination of AR genes in microbial communities, particularly, within the human gut (McInnes et al., 2020). We evaluated whether the Hi-C MAGs portrayed a more complete gene and plasmid content of the species than the WGS (**Figure 6**).

Noteworthy, while only a handful of genes were unique to the WGS-MAGs ($n = 35–76$), the sets of Hi-C-unique genes were as large as $n = 752–1,095$. Decomposition of the latter showed that, although the Hi-C MAGs' contamination was low (<2.5%), their total length was higher than of their WGS counterparts not just due to a higher chromosomal completeness, but also due to the inclusion of plasmid-like contigs – that were underrepresented in the WGS-MAGs ($n \leq 7$ per each). The fact that such plasmid content is not mostly a false-positive inclusion from unrelated taxa is supported by the BLAST search against nr database showing that all classified contigs belonged to *Klebsiella* or related taxa (*Enterobacter*, *Salmonella*, *Escherichia*, *Raoultella*, *Pseudomonas*, *Citrobacter* or

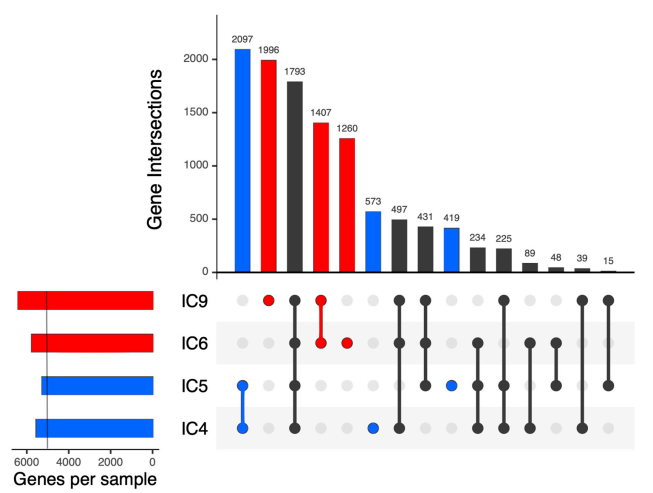


FIGURE 5 | Intersection of the *K. pneumoniae* gene sets across the samples and patients using UpSet plots. Each vertical bar shows the number of genes common across the samples highlighted by dots below the bar. Subject-specific sets are shown in blue and red for patient A and B, respectively. On the left, a light vertical line denotes $n = 5{,}000$ genes.

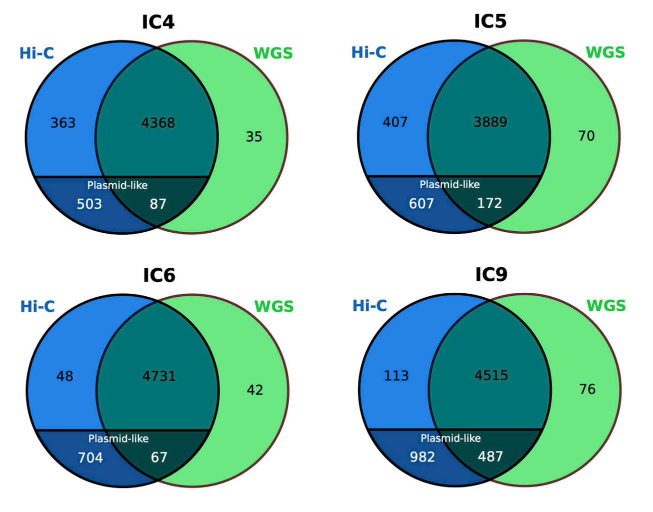


FIGURE 6 | Comparison of the *K. pneumoniae* gene set between the WGS and Hi-C MAGs. The numbers for the plasmid-like contigs of Hi-C MAGs are particularly shown. The Venn diagrams are based on COG annotation.

Acinetobacter – the Gammaproteobacteria order, mostly from *Enterobacteriaceae* family).

We evaluated how Hi-C data can help establish the links of plasmids to their bacterial hosts at the community level. Firstly, identification of potential plasmid-related contigs in the 4 metagenomes suggested an extensive presence of plasmids, with the Proteobacteria and *Enterococci* showing disproportionately high contribution to the plasmid pool compared to the commensal taxa (Additional File 1: **Supplementary Figure 4**). Although the WGS allows the identification of plasmid-like contigs, linking them to bacterial species/MAGs is impaired due to separation from the chromosome and differences in genomic features used for binning like oligonucleotide spectrum. One essential advantage of Hi-C metagenomics compared to the WGS is an ability to link the extrachromosomal content to a MAG *via* a chromosome linking signal. Having assessed the plasmid content of Hi-C MAGs (Additional File 9: **Supplementary Table 9**), we found that the Hi-C MAGs contained a higher number of plasmid contigs than their WGS counterparts (8.8 ± 17.4 vs. 0.94 ± 2.15 across all samples; the cases of highly contaminated WGS-MAGs were not considered here; $p = 0.005761$, Wilcoxon signed rank test with continuity correction, $N = 18$). As a bright example, the *Klebsiella* MAG in the sample IC9 included a plasmid contig of 109,640 bp – representing an almost complete plasmid [according to NCBI nr search, almost perfectly matching a *K. pneumoniae* plasmid previously described for an isolate obtained in Saint Petersburg, Russia (GenBank ID: CP066857.1)] – that was absent in the corresponding WGS-MAG. We further confirmed the circularity of the plasmid by analyzing the assembly graph: the contig ends were overlapping by 55 bp (Additional File 1: **Supplementary Figure 5**).

Antibiotic Resistance Analysis

Using an assembly graph based approach implemented in the GraphAMR pipeline, we evaluated the total resistome of each sample (Additional File 10: **Supplementary Figure 6**). Contrarily to patient A whose semiquantitative AR profiles were remarkably similar between the two time points, the resistome of patient B manifested profound changes in the presence of AR genes (most remarkable were the obtained potential resistance to vancomycin, as shown by occurrence of genes from *van* family, as well as to the tetracyclines suggested by the presence of genes from *tet* family).

We evaluated how Hi-C metagenomics can improve profiling of AR genetic determinants *via* the improved MAG reconstruction. For this analysis, along with the *K. pneumoniae* MAGs, we selected the most abundant Hi-C MAGs (>5%) from each sample (yielding 4–6 MAGs per sample) and identified ARGs in them *via* CARD RGI (see section "Materials and Methods"). Up to 32 AR genes per MAG were detected (including "perfect" and "strict" hits; see Additional File 11: **Supplementary Table 10**). For patient A, most MAGs carried the *adeF* gene; genes conferring resistance to fluoroquinolones and tetracycline were detected. The *Bacteroides* and *Parabacteroides* additionally showed potential resistance to cephamycin.

For patient B, most MAGs included ARGs related to fluoroquinolones and tetracyclines. The *Bacteroides* MAG at the second time point had genes conferring resistance to cephalosporins (*via* CblA-1 gene specific to *B. uniformis*) – but they were not detected in the *Bacteroides* MAG at the first time point. We checked if it was in fact present at the first point, but in a contig that failed to become binned to a MAG. However, it was not found among the unbinned contigs (Additional File 11: **Supplementary Table 10**) – likely having failed to be assembled. Possibly, the observation reflects the cefoperazone/sulbactam treatment of the patient (**Figure 1**).

Following the findings about the higher gene and plasmid contigs counts observed in the Hi-C MAGs compared to their WGS counterparts, we explored the additional value of Hi-C metagenomics in terms of AR. In the example of *Klebsiella*, compared to the WGS-MAGs, the respective Hi-C MAGs included more ARGs – 24–32 vs. 21–22 hits (5–13 vs. 3–4 perfect hits). The Hi-C-specific best ARO (Antibiotic Resistance Ontology) hits (genes, in this context) included, for the patient A, the QnrB1, *dfrA14*, H-NS, QnrS1, OmpA, *msrE* (as opposed to the few WGS-specific ones – QepA2 and TEM-1). For the patient B, the following genes were detected only in the Hi-C MAGs – *mphA*, *dfrA5*, *qacEdelta1*, *sul1*, *sul2*, APH(3′)-VI, NDM-1, QnrS1, *msrE*, *mphE*, BRP(MBL), TEM-1 – while none of the genes were WGS-specific.

Using the Hi-C graph image of resistome, we compared the temporal dynamics of *Klebsiella* resistance potential with the antibiotic regime (see **Figure 1**). For patient A, the QnrB1 and FosA3 genes were unique to the 1st time point, while the QnrS1 and *msrE* – to the second one. At the level of antibiotic classes, unlike the 1st point, the 2nd time point was characterized by the presence of genes conferring resistance to lincosamide, streptogramin, oxazolidinone and pleuromutilin. Noteworthy, both points were characterized by resistance to carbapenems and aminoglycosides – which is in line with the administered meropenem and amikacin, respectively.

Similar analysis for the *Klebsiella* in patient B microbiome showed that 4 ARGs were baseline-unique (*dfrA5*, *qacEdelta1*, *sul1* and FosA5) and 3 (FosA6, TEM-1 and *catI*) – specific for the 2nd point. No differences between the timepoints were observed at the level of antibiotic classes. Interestingly, while the patient was treated with trimethoprim/sulfamethoxazole, no genes conferring resistance to the drug was identified in the *Klebsiella* MAGs – it is in line with the observation that its relative abundance strongly decreased (from 5.5 to 0.6%) suggesting therapy effectiveness.

We assessed how the Hi-C-mediated linking of plasmids to bacterial chromosomes improved capturing of bacterial resistance profiles. To do it, we calculated the proportion of ARGs located on chromosomal contigs for the *Klebsiella* MAGs (Additional File 12: **Supplementary Table 11**). For patient A, most ARGs were located on the chromosome (extrachromosomal 8.3 – 12.5% of all hits and 2–3 out of 5 perfect hit genes). On the contrary, for patient B the extrachromosomal proportion of ARGs was considerably higher (34.4–38.7% of all hits and 10/12–13 of perfect hit genes). Considering the fact that WGS-MAGs almost lacked plasmid-like contigs, for patient B, the resistome would have been considerably underestimated without application of the Hi-C data (and inclusion of plasmid contigs to MAG).

Comparative GraphAMR analysis of the resistome of patient B between the two time points suggested the acquired resistance to vancomycin at the second point (sample IC9) (Additional File 10: **Supplementary Figure 6**) with many ARGs present (*vanA, vanH, vanX, vanR, vanS, vanZ, vanY*). To date, several different types of glycopeptide resistance have been characterized (Arthur et al., 1993); these correspond to specific operons present in the species (Leclercq and Courvalin, 1997). The presence of *vanA* gene suggests the VanA-type resistance, which was the first among the characterized ones and is the most common. This kind of resistance is mediated by transposon Tn1546 or closely related elements. The complete sequence of this ∼10.8 Kbp mobile element was absent in the assembled contigs, and therefore we analyzed the assembly graph neighborhood of ARG matches as reported by GraphAMR to reconstruct the putative structure of VanA operon.

The sequences of genes of interest were located in a tangled (repeat-rich) region on a plasmid and were scattered across 3 edges of the assembly graph (see Additional File 1: **Supplementary Figure 7**). All 7 genes were found in the correct order, however, the topology of the assembly graph and the observed read coverage of the edges suggest that two different variants of Tn1546 transposon are actually present in the sample: one containing the IS1251 mobile element and another one – without this element ('C' and 'A' types of Tn1546 transposons as defined in Wardal et al., 2017). Given the short lengths of the corresponding graph edges (2,122, 2,837 and 1,951 bp), it is not surprising that the assembler was unable to resolve these long repeats and join them into a single contig/scaffold. Exploration of the Hi-C contacts between contigs and MAGs showed that the discovered mobile element in the sample IC9 likely belongs to the *Enterococcus faecium* (normalized link weight between these contigs and *E. faecium* MAG was 0.87 ± 0.84; see section "Materials and Methods" for details on normalization).

Finally, in a complementary MAG-independent network analysis, we evaluated how Hi-C data can be used to detect the bacterial hosts of the ARGs by linking ARG-containing plasmid-like contigs to chromosomal ones. As examples, we used the *Bacteroides cfxA* gene along with the *Klebsiella* OXA and TEM genes – each abundant in sample IC6 and located on plasmid contigs. For each gene, we constructed a network of IC6 contigs linked with Hi-C read pairs around the contig containing the gene (only the contigs > 1,000 bp; the links were normalized as described in section "Materials and Methods"). The obtained environment of the selected ARGs is shown in **Figure 7**. The *cfxA* gene appeared to be linked with the cluster of contigs classified as Bacteroidetes phylum, specifically as *Bacteroides*, in agreement with the existing knowledge. The OXA and TEM genes were both located on the same contig and linked to Proteobacteria chromosomal contigs classified (using Kraken) as belonging to the *Klebsiella* or *Shigella* genera. Interestingly, this contig was not included into any of the MAGs – showing how information about antibiotic resistance undetected under MAG-based approaches can be identified only by using the contig-level network analysis.

Linking Prophages to Bacterial Hosts Using Hi-C

Phages are considered to play important roles in microbial ecology. Previous reports showed that Hi-C data can aid in linking them to their bacterial hosts (Marbouty et al., 2017, 2021; Kent et al., 2020). We investigated this approach on our clinical metagenomes. It started with an observation that the Hi-C MAGs list included the items with very low completeness (according to CheckM) but listed among the most abundant MAGs; this effect was observed for both samples of the patient B (IC6 and IC9). A closer examination of IC9 showed that one of such MAGs is composed of 3 contigs classified as crAssphage; their cumulative length was close to the typical genome length for this phage

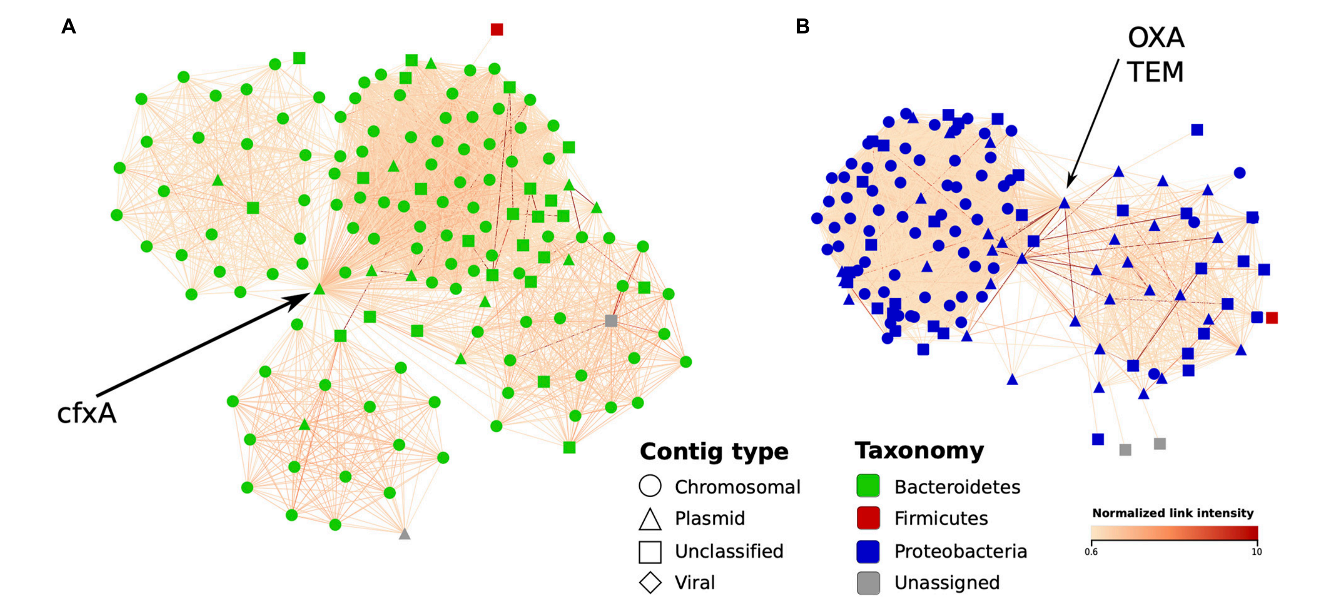


FIGURE 7 | Graph environment of selected antibiotic resistance genes in a Hi-C-based contig network (sample IC6). In each network, vertices correspond to contigs close to a contig containing an AR gene, while the edges – to Hi-C links between the contigs. The layout was constructed using the edge-weighted spring-embedded algorithm. Contig type is shown in shape and bacterial contigs' taxonomy – in color; edge color denotes intensity of normalized Hi-C link. **(A)** The first graph region of interest: a plasmid-like contig carrying *cfxA* gene is linked to a cluster of contigs classified as belonging to *Bacteroides* genus. **(B)** The second region of interest: a plasmid-like contig carrying OXA and TEM genes linked to a cluster of contigs mostly classified as *Klebsiella*.

(98 Kbp) suggesting its high completeness. This finding was in agreement with the results of MiCoP showing crAssphage among the most abundant viral sequences (72 and 88% of total viral abundance for IC6 and IC9, respectively). However, investigation of Hi-C links of the crAssphage MAGs did not show link above the established threshold of 0.6 in normalized Hi-C networks for both samples; noteworthy, in the sample IC6, the only suggestive link (0.4) of crAssphage was with the *Bacteroides dorei* MAG. Considering that relative abundance of *B. dorei* did not change much between the two timepoints, one could speculate that at the first point the crAssphage could be presented as a prophage only in a part of the *B. dorei* population, while at the second point – as free phages.

To get a wider perspective on the phages' hosts, we analyzed viral composition of our samples and evaluated possible "bacteria-virus" associations in each sample, as well as between two time points for each patient. For each sample assembly, viral contigs were identified and annotated (at order or family level). From 1,156 to 1,612 contigs per sample initially defined as viral, the numbers remaining after taxonomic assignment and verification were 333–505; their median length was 3,092–3,233 bp. A network of viral contigs and bacterial MAGs was constructed by considering all Hi-C links between them with normalized weights > 0.6 (**Figure 8** and Additional File 1: **Supplementary Figure 8**). Some viral contigs formed dense clusters with one MAG, while others had strong Hi-C links with several MAGs, thus forming entangled networks. Interestingly, while we detected 107–159 viral contigs per sample each associated with just one MAG, there were many (12–167) contigs per sample that had strong connections to multiple MAGs (Additional File 1: **Supplementary Figure 9**). Binning algorithms conservatively assigned such viral contigs to at most one MAG, while direct assessment of Hi-C links provided the way to discover all possible hosts of a virus.

Furthermore, some viral contigs have a great number of Hi-C links between each other, possibly reflecting low assembly quality, when the algorithm failed to assemble a single viral sequence and left multiple contigs. Patient B has a sparse network at the first time point characterized by a number of distinct clusters – in contrast to the 2nd time point when the number of links increased considerably to form a dense network. This might reflect changes in microbial ecology linked to intensive therapy.

We additionally explored the taxonomic patterns of "bacteria-virus" associations by estimating the Hi-C links between bacterial MAGs and viral contigs across the samples (Additional File 1: **Supplementary Figure 10**). Overall, there were 8 viral families associated with bacteria in at least one of the samples; the number of bacterial families detected as being involved in these links was 25. Among the viruses, the most prevalent connections were represented by *Siphoviridae* and *Myoviridae* families (along with the unclassified families from the *Caudovirales* order). In each sample, each of these three viral groups had links with bacterial hosts – most frequently with the members of *Enterobacteriaceae*, *Oscillospiraceae* and *Desulfovibrionaceae* families. On average, the *Siphoviridae* viruses manifested the highest number of links – in agreement with the fact that it was the most abundant family among the viral contigs. These results show how Hi-C signals help to reveal the taxonomic complexity of interactions between phages and their hosts in the human gut.

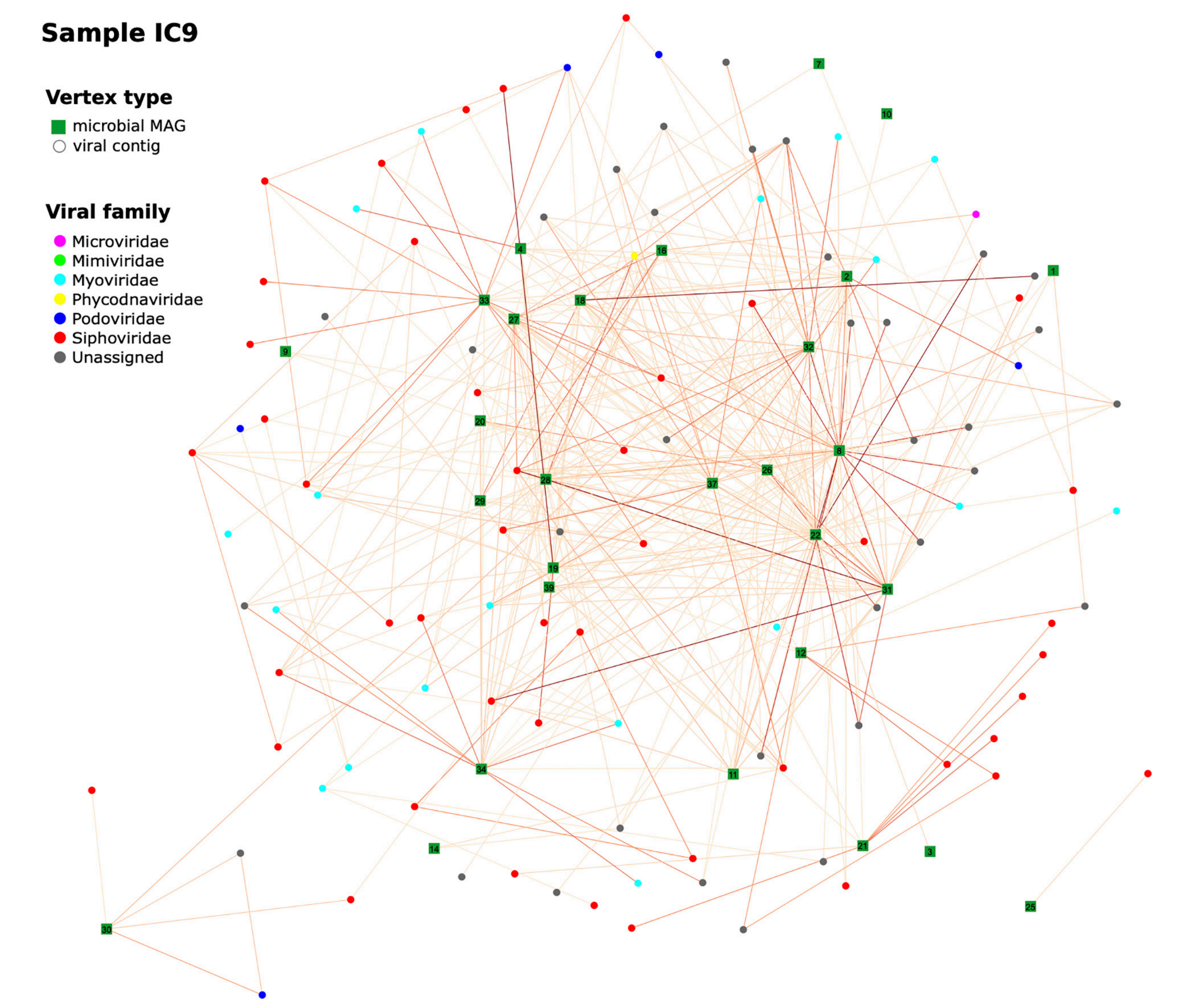


FIGURE 8 | Hi-C-based "virus-bacterial host" networks. The network for the sample IC9 is shown (see Additional File 1: **Supplementary Figure 6** for the other samples). Edge-weighted Spring-Embedded algorithm was applied to generate the layout. Vertex shape corresponds to type (MAG or viral contig). The contigs forming each large microbial MAG were combined and shown as a square labeled with MAG ID. Each contig classified as viral was shown as a circle (particularly, if a viral contig had been assigned to a MAG, here it was extracted and drawn as a circle). Vertex color corresponds to viral taxonomy annotation obtained *via* DemoVir. The intensity of edge colors corresponds to the maximum of weight (normalized Hi-C linkage signal) between the contig and all contigs of the MAG. The MAG IDs are as listed in Additional File 3: **Supplementary Table 3**.

To compare these findings with a complementary approach, we looked for links between viral contigs and high-quality MAGs using the VirMatcher tool based on multiple criteria (see section "Materials and Methods"). Overall, the number of the discovered associations was 22–76 per sample (Additional File 13: **Supplementary Table 12**). When compared with the results of our Hi-C analysis, the overlap was 8–30 links confirmed by both methods (in this way, 3.0–6.1% of the Hi-C findings were supported *via* VirMatcher). There could be various reasons for incomplete confirmation. Besides methodological ones like possible false-positive detections of Hi-C network approach and insufficiently complete assembly, the bacteria might lack CRISPR corresponding to a virus but still have alternative methods of defense against it, to name a few – chemical defense, preventing adsorption, restriction–modification and related defenses, and Argonaute proteins (Hampton et al., 2020).

DISCUSSION

In one of the first Hi-C-aided clinical metagenomic studies presented here, we applied Hi-C metagenomics for deeper exploration of the gut microbiome in chronically critically ill patients. On the example of most abundant opportunists we show how this approach can be beneficial in the clinical context.

Augmentation of WGS sequences with the microbiome-wide chromosome conformation capture data (Hi-C) during binning of contigs resulted in better bacterial genome reconstructions – higher proportion of binned contigs, higher number of quality MAGs and contamination lower by 1–2 orders of magnitude. Furthermore, the novel hicSPAdes algorithm proved to perform better than bin3c, a representative state-of-art Hi-C-based binner. The key difference between hicSPAdes and bin3c is that the latter operates on assembled contigs, while the former uses both contigs and assembly graph. The additional use of topology of the assembly graph improves the binning results as it circumvents the Hi-C data coverage gaps, allows better resolution of repetitive contigs, etc. As a result, the bins obtained by hicSPAdes are more complete and pure. The hicSPAdes is supplemented by the BinSPreader stage that further refines the binning as it allows for assignment of contigs to multiple bins at the same time and splitting the input reads for subsequent reassembly of individual MAGs (this procedure may further improve the contiguity of MAGs due to lesser influence of interspecies repeats).

Comparison of the community structures across the analyzed samples highlighted *Klebsiella pneumoniae* as an omnipresent opportunist which can be used as an example to explore the opportunities of Hi-C metagenomics. This species is a major cause of hospital-acquired infections world-wide including pneumonia, urogenital tract infections (UTIs) and bloodstream infections, especially in immunocompromised individuals, and represents a substantial healthcare burden (Martin and Bachman, 2018). Its morbidity potential is amplified by extensive virulence potential and multi-drug resistance encoded in its open accessory genome, much of which is carried on plasmids. We demonstrated how the consideration of Hi-C data during MAG reconstruction improves the capture of *K. pneumoniae* plasmid content, particularly, the antibiotic resistance genes. It allowed us to detect important virulence factor genes absent in the WGS profiles and, moreover, to identify the inter-individual differences in VF content which was not revealed by conventional metagenomics. Hi-C metagenomics looks especially promising for improving the reconstruction of the mobile genetic elements within genomes that represent problematic regions for assembly and binning. Better reconstruction of MAG improved the accuracy of the downstream comparative genomic analysis; besides the steps described in the manuscripts, it has implications for SNP/indels analysis, phylogenetics and so on. The approach can be applied to other opportunist taxa actively involved in HGT and notorious for their virulent and multidrug-resistant members like *Enterococcus* and *Escherichia*.

There were interesting observations among the taxonomic compositions that might have clinical significance for the critically ill patients. In 3 of 4 samples, we detected a high abundance of *Cloacibacillus* – *C. porcorum* or related – an amino acid degrading microorganism capable of using mucin as a sole carbon source (Looft et al., 2013). The species is a potential gut beneficiary of chronic critical illness linked to muscle loss (cachexia) and malabsorption. Particularly, in patient B, at the 2nd time point it might have replaced another – commensal – mucose-dwelling species *Akkermansia muciniphila* abundant at 1st point. Three of the 4 samples included an abundant MAG classified as OEMR01, a member of the *Erysipelotrichaceae* family. The links of *Erysipelotrichaceae* members to host health are yet to be elucidated; they appear to be highly immunogenic and can thrive after treatment with broad-spectrum antibiotics (Zhao et al., 2013; Kaakoush, 2015). According to our previous 16S rRNA sequencing study, the *Erysipelotrichaceae* was enriched in the gut microbiome of CCI patients compared to the patients in acute critical state (Chernevskaya et al., 2020).

In a clade-specific marker analysis, we identified fungal sequences in each sample. Fungi can represent significant health risks for critically ill patients. As the coverage was low, we did not recover fungal MAGs, not to mention the sufficient Hi-C signal (additional experimental enrichment of the fungal fraction would be required). We anticipate that in such cases, involvement of Hi-C metagenomics to bin fungal genomes consisting of multiple chromosomes will be indispensable. One of the interesting results of the study is the dominance of *Enterocytozoon bieneusi* in the composition of fungiome, an obligate intracellular parasite infecting intestinal cells agent of intestinal microsporidiosis that can manifest as diarrhea (Weiss and Becnel, 2014). The condition can be life-threatening in immunocompromised patients, particularly in the chronically critically ill group.

Overall, the Hi-C-assisted MAG reconstruction performed well for the sufficiently covered microbial genomes. Recovery of low-abundant microorganisms would require higher targeted sequencing coverage. Noteworthy, as the ICU patients often manifest low alpha-diversity (intestinal domination of a single species as an extreme case), the chance of obtaining good-quality genome reconstructions is higher than for healthy subjects hosting more diverse communities. We found that not all high-covered taxa produced good-quality MAGs. This might be related to the variability of GC content. One of the possible experimental solutions would be to use multiple restriction enzymes during the Hi-C library preparation based on the sequence analysis of major expected genomes (Magnitov et al., 2020).

Completeness, the central measure of prokaryotic MAG quality, is commonly based on evaluation of chromosomal single-copy core genes and thus does not take plasmids into account (Parks et al., 2015). Meanwhile, their gene content can drastically affect the bacterial host phenotype, which is especially important for the clinically relevant gut microorganisms. Hi-C metagenomics renders the plasmid content of species detectable and allows to come up with a concept of "plasmid completeness" of microbial genomes reconstructed from metagenomes.

A crucial domain of microbial phenotypes in the clinical context is their drug resistance. The Hi-C data allowed improving resistome profiling – as seen even at the level of MAGs. Although chronic critical illness following severe non-traumatic brain damage was common in these patients, they showed different clinical – as well as microbiome – trajectories. The non-survived patient A had been given antibiotics for a long time prior to the first time point and her therapy was quite constant between the timepoints. At the level of her microbiome, it was reflected by similar species-level composition at the two points and resistome – the latter being comparable by both total ARG relative abundance and presence patterns. On the other hand, the microbiome of patient B (ultimately recovered) who started antibiotic administration at the 1st time point was characterized by strong changes in taxonomic composition with quantitative and qualitative alterations of the resistome. As the pilot sample size was small and the set of prescribed antibiotics varied between the patients, we cannot claim significant effects of the therapy on gut resistome. Various administered drugs showed diverse patterns. For example, for patient A, at baseline, potential resistance to some discontinued drugs increased (possibly reflecting the recovery of a low-abundant resistant population), while for some it decreased (might be removed by negative selection).

The results of ARG prediction even in average-complexity metagenomes (such as the human gut) could be significantly affected by fragmented assemblies. We demonstrated that the use of assembly graph-based approaches is far superior in terms of recovery of more complete ARG sequences even from fragmented metagenome assemblies. Specialized pipelines such as GraphAMR could be used to improve the current approaches of ARG prediction of metagenomic assemblies. Hi-C data could be used to further validate and confirm the results obtained. One essential problem here is that during the assembly, an ARG sequence present in multiple species is likely to be included – flattened into a part of a single contig – into a single MAG. Our results suggest that for deeper resistome profiling using Hi-C, it is promising to operate directly on assembly graphs – prior to formation of contigs and binning of them into MAGs.

Another entity in the gut microbiome that is highlighted by the Hi-C metagenomics are phages. Phages are considered to contribute considerably to the regulation of gut microbial communities (Sutton and Hill, 2019). Identification of their bacterial hosts can help elucidate the precise mechanisms of their contribution. One of the most studied phage families are crAss-like phages; crAssphage's host has been identified to be *Bacteroides intestinalis* (Shkoporov et al., 2018). Previously, it was shown for the same population as the present study (Russian) that the crAssphage reads can represent as much as 24% of the stool metagenome (Yarygin et al., 2017). Recent study in healthy subjects showed using the Hi-C metagenomics how phages of this group can be linked to various species within the *Bacteroides* genus (Marbouty et al., 2021). In our study, although we discovered high levels of crAssphage persisting in one of the patients between the time points, there were no strong links to any bacterial MAGs. The fact that there was a slight contact to *Bacteroides dorei* at the first point only suggests underlying dynamics of proportions between prophages and free phages. After expanding our analysis from this providential occurrence to a global analysis of "virus-bacterial host" network using Hi-C, we discovered the presence of viral contig hubs linked to multiple hosts. Although this could partially be due to misassemblies, such results may hint at possible promiscuity of phages. This can have implications for transmission of ARGs and virulence factors determinants across diverse gut species in immunocompromised patients (however, we have not detected ARGs in viral contigs in our data).

One of the challenges in the present study was to discern signal from noise in Hi-C metagenomic data. It is possible to determine a proper threshold *via* additional experiments on defined bacterial consortia with plasmids, preferably those of high diversity comparable to human gut. In the absence of such opportunities, we determined the threshold by assessing the inter-intra-MAG links distributions. In our case, the separation of distributions choice was visually similar between the 4 samples and the false discovery rate of detecting an inter-MAG was quite low (from 0.0082 for IC6 sample to 0.0351 for IC5, see **Supplementary Figure 2B**). The specific threshold value is likely to vary for new datasets or under experimental protocol modifications. Therefore, it is recommended to evaluate such distributions for each particular dataset.

Another limitation is a small sample size – that did not provide an opportunity to assess statistical significance in some analyses. However, our study did not set an objective of comparing the two patients with each other, but we had rather initially selected the most interesting representative examples of CCI patients in order to illustrate the broad possibilities of the clinical Hi-C metagenomics as a method. In connection to the specific clinical group (the critically ill patients), the analysis of statistical power and sufficient sample size face a heterogeneity challenge: in the ICU, the treatment (including the choice of antibiotics) based on individual clinical status and dynamics strongly varies across the subjects. Therefore, the inter-subject variability of clinical factors is much higher than for typical major diseases linked to microbiome composition in metagenome-wide association studies (like inflammatory bowel disease or type 2 diabetes). It follows that for a strict statistical analysis – with proper adjustment for the confounding factors – and considering the inherent high dimensionality and multimodality of microbiome data, the required sample size might be very large. In our previous survey of CCI patients' gut microbiome (Chernevskaya et al., 2021), even among 44 patients at the group level we did not observe a prevalent antibiotic therapy pattern – the individual treatments were highly variable. Nevertheless, despite different diagnoses, the chronically critically ill patients convergently acquire the same features of the clinical course, with profound changes in the gut microbiome. Thereby, the dataset in this study allowed us to demonstrate how Hi-C metagenomics can be expedient in the context of clinical metagenomics. The technique is yet to become affordable for wide application. However, further analysis of larger cohorts might provide the basis for developing simpler targeted and cost-effective methods like 3-C for specific clinical aims.

In the ICU microbiome research, the technique can be readily applicable to analysis of other body sites, as well as for hospital surfaces – that can serve as media of pathogens transmission. It is also relevant to the COVID-19 pandemics: it is estimated that not less than ~7% of COVID-19 patients develop bacterial co-infection and most lethal outcomes in the ICU are ultimately determined by this factor. As the gut microbiome is an important reservoir of opportunistic infectious agents causing such invasions, its virulence and drug resistance potential should be explored in detail.

CONCLUSION

Hi-C metagenomics is a promising tool for analyzing clinical microbiome samples. Compared to conventional metagenomics, it provides reconstructed microbial genomes of higher completeness and lower contamination. In the context of critical care, the method coupled with specialized algorithms improves the precision of profiling antibiotic resistance and virulence potential of opportunist gut taxa, as well as the tracking of mobile genetic elements dynamics. The findings can help optimize the treatment schemes and understand mechanisms of pathogenesis in the ICU.

DATA AVAILABILITY STATEMENT

The WGS and Hi-C sequencing data are available in the European Nucleotide Archive (ENA) repository under the accession number PRJNA718195. The data processing scripts are available at https://bitbucket.org/ibg_super/hicicuscripts/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Local Ethics Committee of Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology. Formal consent for participation in this study was obtained from the legal representative of each patient.

AUTHOR CONTRIBUTIONS

AT, SU, SR, and EC conceived the study design, experiments and analyses. EC and NB managed sample collection. PV supervised by SU performed sample preparation. VI, AI, IT, VU, and DS developed the software and analyzed the data. VI, AT, and AK interpreted the analyses. AT, SU, AK, and VU coordinated the project. VI, AT, EC, AK, and AI wrote the manuscript with critical revision performed by SU, SR, VU, NB, and PV. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Russian Science Foundation [grants (19-74-10092) to VI, EC, PV, and AT and (19-14-00172) to AK and IT]. AT was supported by the grant (075-15-2019-1661) from the Ministry of Science and Higher Education of the Russian Federation.

ACKNOWLEDGMENTS

We thank Ivan Elizarov and Alexandr Merkel for aid in sample preparation, and the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, IGB RAS – for providing computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.770323/full#supplementary-material

REFERENCES

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

Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P. A. (2020). Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* 36, 4126–4129. doi: 10.1093/bioinformatics/btaa490

Arndt, D., Marcu, A., Liang, Y., and Wishart, D. S. (2019). PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. *Brief. Bioinform.* 20, 1560–1567. doi: 10.1093/bib/bbx121

Arthur, M., Molinas, C., Depardieu, F., and Courvalin, P. (1993). Characterization of Tn1546, a Tn3-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. *J. Bacteriol.* 175, 117–127. doi: 10.1128/jb.175.1.117-127.1993

Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R., and Marbouty, M. (2019). MetaTOR: a computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (meta3C) libraries. *Front. Genet.* 10:753. doi: 10.3389/fgene.2019.00753

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., et al. (2022). Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. *Nat. Biotechnol.* doi: 10.1038/s41587-021-01130-z

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

Chernevskaya, E., Beloborodova, N., Klimenko, N., Pautova, A., Shilkin, D., Gusarov, V., et al. (2020). Serum and fecal profiles of aromatic microbial metabolites reflect gut microbiota disruption in critically ill patients: a prospective observational pilot study. *Crit. Care* 24:312. doi: 10.1186/s13054-020-03031-0

Chernevskaya, E., Klimenko, N., Pautova, A., Buyakova, I., Tyakht, A., and Beloborodova, N. (2021). Host-microbiome interactions mediated by phenolic metabolites in chronically critically ill patients. *Metabolites* 11:122. doi: 10.3390/metabo11020122

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an r package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Cox, C. E. (2012). Persistent systemic inflammation in chronic critical illness. *Respir. Care* 57, 859–64; discussion 864–66. doi: 10.4187/respcare.01719

DeMaere, M., and Darling, A. (2019). bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol.* 20:46. doi: 10.1186/s13059-019-1643-1

DeMaere, M. Z., and Darling, A. E. (2018). Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies. *GigaScience* 7, 1–12. doi: 10.1093/gigascience/gix103

DeMaere, M. Z., and Darling, A. E. (2021). qc3C: reference-free quality control for Hi-C sequencing data. *PLoS Comput. Biol.* 17:e1008839. doi: 10.1101/2021.02.24.432586

DeMaere, M. Z., Liu, M. Y. Z., Lin, E., Djordjevic, S. P., Charles, I. G., Worden, P., et al. (2020). Metagenomic Hi-C of a healthy human fecal microbiome transplant donor. *Microbiol. Resour. Announc.* 9:e01523-19. doi: 10.1128/MRA.01523-19

Dufresne, K., Saulnier-Bellemare, J., and Daigle, F. (2018). Functional analysis of the chaperone-usher fimbrial gene clusters of *Salmonella enterica* serovar typhi. *Front. Cell. Infect. Microbiol.* 8:26. doi: 10.3389/fcimb.2018.00026

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'Omics data. *PeerJ* 3, e1319. doi: 10.7717/peerj.1319

Foster, J. A. (2016). Gut microbiome and behavior: focus on neuroimmune interactions. *Int. Rev. Neurobiol.* 131, 49–65. doi: 10.1016/bs.irn.2016.07.005

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., and Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with insilicoseq. *Bioinformatics* 35, 521–522. doi: 10.1093/bioinformatics/bty630

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., and Sullivan, M. B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* 28, 724–740.e8. doi: 10.1016/j.chom.2020.08.003

Hampton, H. G., Watson, B. N., and Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature* 577, 327–336. doi: 10.1038/s41586-019-1894-8

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

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

Kaakoush, N. O. (2015). Insights into the role of *Erysipelotrichaceae* in the human host. *Front. Cell. Infect. Microbiol.* 5:84. doi: 10.3389/fcimb.2015.00084

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., et al. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359. doi: 10.7717/peerj.7359

Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J., and Brito, I. L. (2020). Widespread Transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* 11, 4379. doi: 10.1038/s41467-020-18164-7

Klimenko, N. S., Tyakht, A. V., Popenko, A. S., Vasiliev, A. S., Altukhov, I. A., Ischenko, D. S., et al. (2018). Microbiome responses to an uncontrolled short-term diet intervention in the frame of the citizen science project. *Nutrients* 10:576. doi: 10.3390/nu10050576

Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* 46:e35. doi: 10.1093/nar/gkx1321

LaPierre, N., Mangul, S., Alser, M., Mandric, I., Wu, N. C., Koslicki, D., et al. (2019). MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. *BMC Genomics* 20(Suppl. 5):423. doi: 10.1186/s12864-019-5699-9

Leclercq, R., and Courvalin, P. (1997). Resistance to glycopeptides in enterococci. *Clin. Infect. Dis* 24, 545–54; quiz555–6. doi: 10.1093/clind/24.4.545

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 4, D687–D692. doi: 10.1093/nar/gky1080

Looft, T., Levine, U. Y., and Stanton, T. B. (2013). *Cloacibacillus porcorum* sp. nov., a mucin-degrading bacterium from the swine intestinal tract and emended description of the genus *Cloacibacillus*. Int. J. Syst. Evol. Microbiol. 63(Pt 6), 1960–1966. doi: 10.1099/ijs.0.044719-0

Magnitov, M. D., Kuznetsova, V. S., Ulianov, S. V., Razin, S. V., and Tyakht, A. V. (2020). Benchmark of software tools for prokaryotic chromosomal interaction domain identification. *Bioinformatics* 36, 4560–4567. doi: 10.1093/bioinformatics/btaa555

Marbouty, M., Baudry, L., Cournac, A., and Koszul, R. (2017). Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* 3:e1602105. doi: 10.1126/sciadv.1602105

Marbouty, M., Thierry, A., Millot, G. A., and Koszul, R. (2021). MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. *eLife* 10:e60608. doi: 10.7554/eLife.60608

Martin, R. M., and Bachman, M. A. (2018). Colonization, infection, and the accessory genome of *Klebsiella pneumoniae*. *Front. Cell. Infect. Microbiol.* 8:4. doi: 10.3389/fcimb.2018.00004

McInnes, R. S., McCallum, G. E., Lamberte, L. E., and van Schaik, W. (2020). Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Curr. Opin. Microbiol.* 53, 35–43. doi: 10.1016/j.mib.2020.02.002

Nelson, J. E., Cox, C. E., Hope, A. A., and Carson, S. S. (2010). Chronic critical illness. *Am. J. Respir. Crit. Care Med.* 182, 446–454.

Nierman, D. M., and Nelson, J. E. (2002). Chronic critical illness. *Crit. Care Clin.* 18, xi–xii. doi: 10.1016/s0749-0704(02)00017-9

Parfenov, A. L., Petrova, M. V. I, Pichugina, M., and Luginina, E. V. (2020). Comorbidity development in patients with severe brain injury resulting in chronic critical condition (Review). *Gen. Reanimatol.* 16, 72–89. doi: 10.15360/1813-9779-2020-4-72-89

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

Press, M. O., Wiser, A. H., Kronenberg, Z. N., Langford, K. W., Shakya, M., Lo, C. C., et al. (2017). Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv* [Preprint] doi: 10.1101/198713

Prjibelski, A. D., Antipov, D., Meleshko, D., Lapidus, A. L., and Korobeynikov, A. I (2020). Using SPAdes *de novo* assembler. *Curr. Protoc. Bioinform.* 70:e102. doi: 10.1002/cpbi.102

Shafranskaya, D., Chori, A., and Korobeynikov, A. (2021). Graph-based approaches significantly improve the recovery of antibiotic resistance genes from complex metagenomic datasets. *Front. Microbiol.* 12:714836. doi: 10.3389/fmicb.2021.714836

Shkoporov, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., and Ross, R. P. (2018). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* 9:4781. doi: 10.1038/s41467-018-07225-7

Stalder, T., Press, M. O., Sullivan, S., Liachko, I., and Top, E. M. (2019). Linking the resistome and plasmidome to the microbiome. *ISME J.* 13, 2437–2446. doi: 10.1038/s41396-019-0446-4

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* 9:870. doi: 10.1038/s41467-018-03317-6

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinform.* 47, 8.13.1–24. doi: 10.1002/0471250953.bi0813s47

Sutton, T. D. S., and Hill, C. (2019). Gut bacteriophage: current understanding and challenges. *Front. Endocrinol.* 10:784. doi: 10.3389/fendo.2019.00784

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Tyakht, A. V., Kostryukova, E. S., Popenko, A. S., Belenikin, M. S., Pavlenko, A. V., Larin, A. K., et al. (2013). Human gut microbiota community structures in urban and rural populations in Russia. *Nat. Commun.* 4:2469. doi: 10.1038/ncomms3469

Volokh, O., Klimenko, N., Berezhnaya, Y., Tyakht, A., Nesterova, P., Popenko, A., et al. (2019). Human gut microbiome response induced by fermented dairy product intake in healthy volunteers. *Nutrients* 11:547. doi: 10.3390/nu11030547

Wardal, E., Kuch, A., Gawryszewska, I., Żabicka, D., Hryniewicz, W., and Sadowy, E. (2017). Diversity of plasmids and Tn1546-Type transposons among vana *Enterococcus faecium* in Poland. *Eur. J. Clin. Microbiol. Infect. Dis.* 36, 313–328. doi: 10.1007/s10096-016-2804-8

Weiss, L. M., and Becnel, J. J. (2014). *Microsporidia: Pathogens of Opportunity*. Hoboken, NJ: John Wiley & Sons.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46.

Yarygin, K., Tyakht, A., Larin, A., Kostryukova, E., Kolchenko, S., Bitner, V., et al. (2017). Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biological hypotheses. *PLoS One* 12:e0176154. doi: 10.1371/journal.pone.0176154

Zhao, Y., Wu, J., Li, J. V., Zhou, N. Y., Tang, H., and Wang, Y. (2013). Gut microbiota composition modifies fecal metabolic profiles in mice. *J. Proteome Res.* 12, 2987–2999. doi: 10.1021/pr400263n

Conflict of Interest: AT was an employee of Knomics LLC and Atlas Biomed Group companies at the time of the study. Sequencing was performed by Knomics LLC. The companies were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership